A NOVEL HIGH-IMPACT PHENOTYPE SELECTION METHOD USING A HYBRID OF AUC & HMM

BY

IMRAN UL HAQ

A Thesis Presented to the

DEANSHIP OF GRADUATE STUDIES

KING FAHD UNIVERSITY OF PETROLEUM & MINERALS

DHAHRAN, SAUDI ARABIA

In Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

In

COMPUTER SCIENCE

DEC 2012

KING FAHD UNIVERSITY OF PETROLEUM & MINERALS
DHAHRAN 31261, SAUDI ARABIA

DEANSHIP OF GRADUATE STUDIES

This thesis, written by **IMRAN UL HAQ** under the direction of his thesis adviser and approved by his thesis committee, has been presented to and accepted by the Dean of Graduate Studies, in partial fulfillment of the requirements for the degree of **MASTER OF SCIENCE IN COMPUTER SCIENCE**.

Thesis Committee

Dr. Moataz Ahmed (Chairman)

Dr. Md. Rafiul Hassan (Adviser)

Dr. Mohammad Alshayeb (Member)

Dr. Mahmood Niazi (Member)

Dr. Sajjad Mahmood (Member)

Dr. Adel F. Ahmed
Department Chairman

Dr. Salam A. Zummo
Dean of Graduate Studies

30/1/13
Date

Dedicated to my parents for their love, support and guidance

ACKNOWLEDGMENTS

I would like to start by thanking Allah Almighty for providing me the opportunity to get a Masters degree from a reputable university like King Fahd University of Petroleum and Minerals (KFUPM). I am gratefully indebted to KFUPM for supporting me throughout my Master studies through their "research assistant" program for international students.

I would like to express my gratitude to my advisor Dr. Md. Rafiul Hassan for his guidance and unconditional support throughout my MS thesis. He taught me everything about doing research, analyzing the results and writing a scientific research paper. I am thankful to my thesis committee chairman Dr. Moataz Ahmed as well as the committee members Dr. Mohammad Alshayeb, Dr. Mahmood Niazi and Dr. Sajjad Mahmood for providing critical feedback on the thesis. I remain grateful to the chairman of the ICS department Dr. Adel Ahmed for his continuous support and guidance throughout my Master studies. His challenging tasks allowed me to learn Microsoft SharePoint and different Adobe tools by developing useful products for the betterment of the department. I am also thankful to Dr. Emad Ramadan who helped me a lot by guiding me in the analysis of the biological aspects of this thesis.

Finally, I am in eternal debt to my wonderful parents for their never-ending love

and continuous support in every aspect of my life.

I would like to acknowledge that without the support and guidance of the above mentioned people, this research would not have been possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ACTB	actin, beta
ACTR1A	ARP1 actin-related protein 1 homolog A
AI	artificial intelligence
AIC	akaike information criterion
ALL	acute myelogenous leukemia
AML	acute lymphoblastic leukemia
ANN	artificial neural network
AUC	area under the curve
BIC	bayesian information criterion
BRCA1	breast cancer 1, early onset
BRCA2	breast cancer 2, early onset
cDNA	complementary DNA
CST3	cystatin C
DAVID	database for annotation, visualization, and integrated discovery
DNA	deoxyribonucleic acid

FN	false negative
FP	false positive
GA	genetic algorithm
GB	gigabyte
GEO	gene expression omnibus
GMT	gene matrix transposed
GPX4	glutathione peroxidase 4
GSEA	gene set enrichment analysis
HADHA	hydroxyacyl-CoA dehydrogenase/3-ketoacyl-CoA thiolase/enoyl-CoA hydratase
HELLP	hemolysis (H), elevated liver enzymes (EL), and low platelet count (LP)
HMM	hidden Markov model
ISCU	iron-sulfur cluster scaffold homolog
LOOCV	leave-one-out cross validation
LSD	least significant difference
MATLAB	matrix laboratory
MCMC	Markov chain monte carlo
Mdl	minimum description length
MDMR	minimum distance to model ranking

MPLS	multivariate partial least squares
MPM	malignant pleural mesothelioma
MPT	multivariate permutation test
MSE	mean square error
PCA	principal component analysis
PD	polychotomous discrimination
QDA	quadratic discriminant analysis
RAM	random access memory
RFE	recursive feature elimination
RFGS	random forrest gene selection
ROC	receiver operating characteristic
SAM	significant analysis of microarray
SNR	signal-to-noise ratio
SOM	self-organizing map
SVM	support vector machine
TN	true negative
TNoM	threshold number of misclassification
TOB1	transducer of ERBB2, 1
TP	true positive
TRANSFAC	transcription factor

WEPO	weighted punishment on overlap

THESIS ABSTRACT

NAME: Imran ul Haq

TITLE OF STUDY: A novel high-impact phenotype selection method using a hybrid of AUC & HMM

MAJOR FIELD: Computer Science

DATE OF DEGREE: 22 Dec 2012

It is well known that the mutations in BRCA1 or BRCA2 gene can cause the hereditary breast cancer. However, it is a tedious and expensive task to identify the mutant genes that impact breast cancer due to the large number of genes and very small number of samples. Researchers have hypothesized that the genes expressed by these two types of tumors are also distinctive. The number of expressed genes could also be very high compared with the actual number of genes that has impact on the cases of breast cancer. Furthermore, the expressive energy of the subset of genes in place of that of one individual gene at a time can be considered to have a profound influence on the cases of breast cancer. Therefore, the objective of this study is to propose a method to identify a small subset of

high-impact genes that are strongly related to breast cancer.

A combination of a non-parametric supervised and an unsupervised statistical method is introduced to analyze the gene expressions and the distinctive genes among the highly expressed genes are identified. The most important genes are filtered using the area under the curve (AUC) measure. These filtered genes are then used to build a hidden Markov model (HMM) to analyze their inter-relationship and identify the best subset among them. In addition, Protein-Protein interaction network is generated to analyze the pathways of the identified genes and their link with BRCA1 or BRCA2. Transcription Factors are identified and Gene Set Enrichment Analysis (GSEA) is calculated for the identified genes subset and the results are compared with the results mentioned in other cancer literature.

The identified genes are not only statistically significant but also illustrate biological significance related to the disease. These genes are also common among the genes that have been identified by other existing studies and gene pathways/ontology analysis. Moreover, the subset of genes extracted by our method is more compact than those previously investigated. Therefore, most of the genes identified by the hybrid method are known to be strongly related to breast cancer.

ملخص الرسالة

الاسم الكامل : عمران الحق أبصار الحق

عنوان الرسالة : طريقة جديدة لإختيار الأنماط الظاهرية ذات التأثير العالي عن طريق دمج ال HMM و ال AUC

التخص : علوم حاسب آلي

تاريخ الدرجة العلمية : كانون أول 2012

من المعلوم أن الطفرات التي تحصل على الجين BRCA1 و الجين BRCA2 سرطان الثدي الوراثي. تعتبر عملية تحديد و إكتشاف الجينات التي تأثرت بالطفرات عمليه صعبة و مكلفة بسبب عدد الجينات الكبير جدا، و العدد القليل من العينات. يفترض العلماء إختلاف الجينات المتأثره بهذين النوعين من الأورام. عدد الجينات المتأثرة بالأورام قد يكون مرتفعا أيضا مقارنة بالعدد الفعلي للجينات التي تؤدي إلى سرطان الثدي. كما أن، الطاقة المؤثرة الناتجة من مجموعه جزئية من الجينات في مكان معين يمكن أن تؤثر بشكل أكبر من تأثير جين واحد في حالة سرطان الثدي. تهدف هذه الدراسة إلى تطور تقنيه لتحديد مجموعة جزئية واحدة من الجينات ذوات التأثير العالي المرتبطة بقوة بسرطان الثدي.

قمنا في هذه الدراسة بتطوير طرق إحصائية غير مراقبة و طرق إحصائية غير حدودية مراقبة لتحليل تعابير الجينات و الجينات المميزة بين الجينات ذات التأثير العالي. تم تصفية الجينات ذات الأهمية العالية بإستخدام مقياس المساحة تحت المنحنى (AUC). بعد ذلك تم إستخدام الجينات المصفاه في عملية بناء نموذج ماركوف المخفي (HMM) من أجل تحليل علاقاتهم الداخلية و تحديد أفضل مجموعة جزئية بينهم. بالإضافة إلى ذلك، تم بناء شبكة تفاعل بين البروتينات بهدف تحليل مسارات الجينات المختارة و إرتباطهم بالجينات BRCA1 و BRCA2. قمنا بتحديد عوامل

النسخ (Transcription Factors)، بالإضافة إلى حساب Enrichment Analysis (GSEA) Gene Set الخاصة بالمجموعات الجزئية للجينات. ثم قمنا بالمقارنة النتائج مع النتائج الخاصة بأبحاث السرطان الأخرى.

لا ينحصر تأثير الجينات المحددة على الجانب الإحصائي فقط، ولكن تبين أنها تملك تأثيرات حيوية مرتبطة مرض. بالإضافة إلى ذلك تظهر هذه الجينات بين الجينات التي تم أكتشافها في الدراسات الأخرى و مسارات الجينات. تتميز المجموعة الجزئية المستخرجة بإستخدام الطريقة المقترحة في هذه الدراسة، بصغر حجمها مقارنة بنتائج الدراسات الأخرى. أخيرا، تبين أن معظم الجينات التي تم تحديدها معروفة بعلاقتها القوية بسرطان الثدي.

Chapter 1

Introduction

The research presented in this thesis develops and analyzes a hybrid computational method to select important high-impact genes in a high dimensional dataset. This research is based on area under the curve (AUC) and hidden Markov model (HMM). A HMM is a statistical process used to model sequential processes in which a future event depends on the current event. Examples that represent sequential process are the sequence of letters in an alphabet and time series data where the collection of observations on variables created sequentially in time [8].

This chapter is organized as follow. The preliminary information regarding the genomic era, bioinformatics, cancer, breast cancer and why is it necessary to better understand breast cancer is presented in Section 1.1. Section 1.2 presents an overview of the gene selection and answer questions like why is it necessary, what are the prominent approaches and what is still missing?. Section 1.3 presents aim of this thesis. Finally, the organization of this thesis is described in Section 1.4.

1.1 Preliminaries

1.1.1 The genomic era

On 26th June 2000, geneticists announced to the world that they had successfully deciphered the human genome. This discovery opened thousands of new doors for scientific research, offering untold opportunities for better health, longer lives and richer human understanding. This marked the beginning of *the genomic era* and resulted in a massive explosion in the amount of biological information available due to huge advances in the fields of molecular biology and genomics [9].

1.1.2 Bioinformatics

This massive explosion also resulted in the rediscovering of the bioinformatics field. *Bioinformatics* is the application of computer technology to the management and analysis of biological data [9]. The ultimate goal of bioinformatics is to uncover the wealth of biological information hidden in the massive data and obtain a clearer insight into the fundamental biology of organisms. This new knowledge can have profound impact on fields as varied as human health, agriculture, the environment, energy and biotechnology.

1.1.3 Cancer

Cancer is a condition in which abnormal cells divide without control and are able to invade other tissues [10]. These cells can begin to grow and divide abnormally and escape our bodys normal control processes. Cancer does not refer to a single disease but many diseases. There are more than 100 different types of cancer [10],

mostly named for the organ or the type of cell in which they initiate, for example colon cancer, skin cancer and breast cancer.

1.1.4 Breast Cancer

Cancer that starts in the tissues of the breast is known as breast cancer [11]. There are two types of breast cancer; hereditary breast cancer that is associated with inherited gene mutations, and non-hereditary (sporadic) breast cancer that is associated with diagnosis without a prior history of cancer in the family. The two most frequent cancer types worldwide are breast and ovarian cancer among women [12]. Trailing lung cancer, breast cancer is the second most common type of cancer in the world and is the fifth most common cause of death from cancer [12].

1.1.5 What's missing?

The above mentioned statistics highlight the importance of understanding breast cancer in more detail. One way to better understand breast cancer is to investigate the patients that have suffered from it. More specifically, in depth examination of the gene microarrays of patients that suffered from breast cancer can result in new information. The genes present in these gene microarrays can be studied at the individual as well as the group level. The ultimate goal is to successfully determine a list of genes whose product abundance can indicate important differences in cell state, such as healthy or damaged, or also in identification of one particular type of cancer or another. Moreover, among these informative genes, some may play a role in the initiation, progression, or maintenance of the disease.

1.2 Gene selection

In the last decade, gene expression data have become an increasingly popular way of exploring the genotypic influences underlying the poorly understood phenotypes. For example, BRCA1 and BRCA2 are well known human genes wherein the mutation of the genes can be linked to the hereditary breast cancer as well as ovarian cancer [13; 14; 15; 16]. Mutations in BRCA gene are somehow heterogeneous, i.e., one normal and the other is mutated copy considering germ-line i.e., the line (sequence) of germ cells that have genetic material that may be passed to a child. Interestingly, BRCA1 mutation carriers are more susceptible to having breast cancer than that of BRCA2 (the risk is up to 80% [16; 17]). Moreover, there is an increased risk of ovarian cancer to the carriers with BRCA1 mutations [18]. However, the risk of BRCA2 mutations can approach to that of BRCA1 with time [19]. Gene expression data analysis can be useful in identifying a gene subset that influences the varying risk factor in between BRCA1 and BRCA2 mutations.

Approaches to extract information from gene expression data fall into several major categories. The earliest and the most straightforward category consists of looking for the individual genes whose expression difference between the two mutations falls significantly close to one extreme, i.e. BRCA1 or BRCA2. These methods search for genes whose uncharacteristically extreme behavior demonstrates that they are related to the underlying phenotypic difference. Investigators then take this list of exceptional genes and try to explain the link between their differential expression and the different conditions they exhibit.

While search for the significance of individual and distinctively expressed genes has yielded many insights [1, 4, 20], it seems increasingly likely that relevant biological deviations in many cases may be too subtle at a single gene to be detectable by such a method [21]. To address this issue, increasing number of methods take broader look into the gene expression and exploit the advantage of the extensive information captured in the data.

1.3 Aim of this research

The aim of this research is to develop, implement, test a coherent and powerful method for the selection of important and high-impact genes from a gene expression dataset, report the test results and compare them with existing studies. A combination of a non-parametric supervised and an unsupervised statistical method is introduced to analyze the gene expression dataset and the distinctive genes among the highly expressed genes are identified.

The aim of this research is to develop a simple and easy to understand hybrid method using a combination of AUC and HMM intended to be a powerful model for selection of important high-impact genes from a high dimensional gene expression dataset.

1.4 Organization of the thesis

The thesis is organized as follows. Chapter 2 contains the background of the various techniques employed in this research. Chapter 3 discusses the existing techniques for feature selection. Chapter 4 presents the summary of recent and

popular researches that has been done regarding feature selection in high dimensional datasets. In Chapter 5, we discuss our proposed hybrid model in detail. Chapter 6 comprises the results that have been achieved from existing techniques discussed in Chapter 3. In Chapter 7, we describe the experimental setup, the dataset used and results achieved from our proposed hybrid method. Chapter 8 discusses the impact of our work where we compare the results of our work with the existing studies biologically and statistically. In the last chapter i.e., Chapter 9, we conclude this research and discuss the significance of our work.

Chapter 2

Background

This thesis discuss the design and analysis of a hybrid feature selection method which combines AUC and HMM. Therefore, a brief theory on these techniques is presented in this chapter. Furthermore, a brief description of Baum-Welch expectation maximization 22 algorithm is also discussed.

This chapter is organized as follows. Section 2.1 presents the preliminary information that need to be understood before discerning ROC and AUC. The receiver operating characteristic (ROC) curve is discussed in Section 2.2. Section 2.3 briefly describes the theory of area under the ROC curve (AUC). Section 2.4 outlines the HMM model. A brief theory of Baum-Welch expectation maximization [22] algorithm is presented in Section 2.5.

2.1 Preliminaries

2.1.1 Classification

In machine learning, classification is a type of supervised learning i.e. learning where a training set of correctly-identified observations is available.

Formally, classification is the problem of identifying to which set of categories, a new instance belongs based on a training set of data containing instances whose class is known.

2.1.2 Binary Classification

In binary classification, only two classes are involved. Formally, each instance I is mapped to one element of the set of positive and negative class labels i.e., {+,-} or {p,n} [23].

2.1.3 Classification model

A classification model, also called classifier, maps instances to predicted classes [23].

In order to distinguish between the actual class and the predicted class, different labels can be used for the class predictions produced by a model like {Y,N}.

2.1.4 Binary Classification Outcomes

The four possible outcomes from a binary classifier and an instance are as follows.

true positive: If the instance is positive and it is correctly classified as positive

false negative: If the instance is positive and it is incorrectly classified as negative

true negative: If the instance is negative and it is correctly classified as negative

false positive: If the instance is negative and it is incorrectly classified as positive

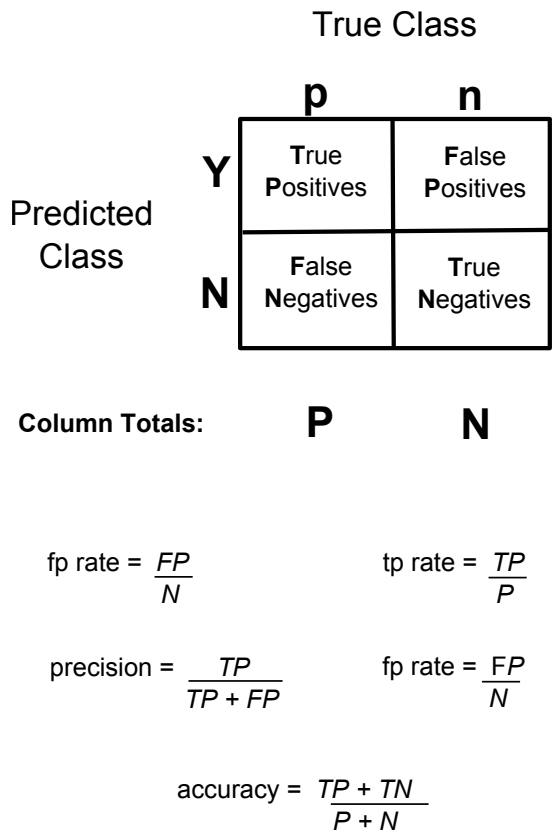

Figure 2.1: Confusion matrix and common performance metrics calculated from it

Let us discuss an appropriate example from a real-world problem in order to better understanding the above mentioned terms. Consider a diagnostic test that tries to determine whether a person has a certain disease or not. A false positive in this example would be when the person tests positive, but he does not have the disease. Conversely, a false negative occurs when the person tests negative, suggesting he is healthy, when he actually have the disease. The other two simpler terms are true positive, test indicate the person having the disease, and true negative the test indicate the person not having the disease.

A 2x2 confusion matrix, also called a contingency table, can be constructed

representing the dispositions of the set of instances from a classifier and a set of instances. Figure 2.1 shows a confusion matrix and equations of several common metrics that can be calculated from it.

2.1.5 Markov Process

A Markov process is a stochastic process where the future event depends on the immediate preceding event. Hence, a Markov process assumes that the probability of the occurrence of the next event will solely depend on the occurrence of the current event.

Figure 2.2: A sequence of letters: A, B and C

Let us understand this with a simple example of a sequence of letters shown in Figure 2.2. In this sequence, the total number of changes, as the next step value in the sequence, from A to A is 2, from A to B is 2, from A to C is 1; from B to A is 2, from B to B is 1, from B to C is 1; from C to A is 1, from C to B is 1 and from C to C is 1. The conditional probabilities for the given sequence are:

$$\mathbf{Pr(A|A) = 2/13;\ Pr(A|B) = 2/13;\ Pr(A|C) = 2/13}$$

$$\mathbf{Pr(B|A) = 2/13;\ Pr(B|B) = 1/13;\ Pr(B|C) = 1/13}$$

$$\mathbf{Pr(C|A) = 1/13;\ Pr(C|B) = 1/13;\ Pr(C|C) = 1/13}$$

The sequence of letters shown in Figure 2.2 can also be illustrated using a directed graph as shown in Figure 2.3 where each node represents a state from which a specific symbol is emitted. For instance, the node labeled as A emits the

symbol A. The lines connecting two nodes represent the probability of changing states; i.e., the state transition probabilities are represented by each of the edges connecting two states.

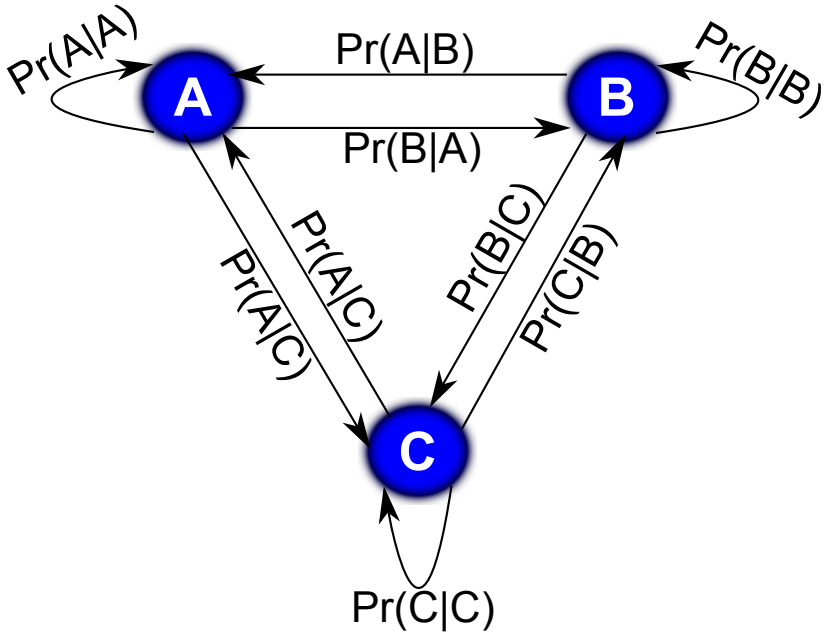

Figure 2.3: The Markov process of the given sequence illustrated by a directed graph

2.2 The receiver operating characteristic (ROC) curve

An ROC curve measure can be used to evaluate the discriminative performance of binary classifiers in machine learning. By varying the discrimination threshold for a binary classifier, the ROC curve measure can be obtained by plotting the curve of true positive rate (sensitivity) versus false positive rate (1 - specificity). When the ROC curve matches with the upper left corner of the ROC space, the best performance would be achieved as this would yield 100% sensitivity and 100% specificity. Moreover, the closer the ROC curve is to the upper part of the ROC space, the better the performance of the classifier.

Figure 2.4 shows ROC curves for 3 different predictors whereas the dotted line shows the line that denotes the average AUC.

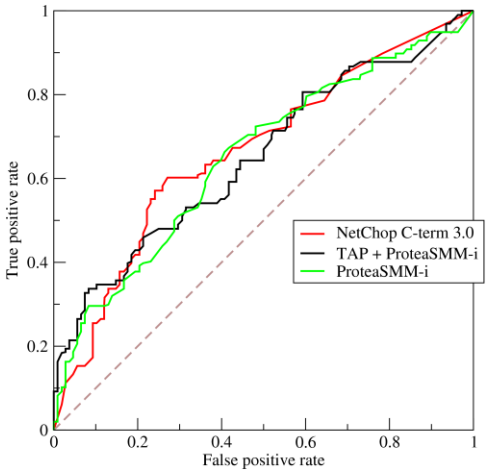

Figure 2.4: ROC curves for 3 different predictors

2.3 Area under the ROC curve (AUC)

Since an ROC curve is a two-dimensional depiction of classifier performance, and in order to compare classifiers we might need to reduce ROC performance to a single scalar value that represents the expected performance [23]. A common method is to calculate the area under the ROC curve, abbreviated AUC. The value of any AUC will always lie between 0 and 1 since the AUC calculates the portion of the area of the unit square [(23)]. However, since random guessing produces the diagonal line between (0, 0) and (1, 1) with an area of 0.5, no realistic classifier should have an AUC less than 0.5. Hence, an AUC value close to 1 indicates better performance [24]. Figure 2.5 shows the areas under two ROC curves, A and B where classifier B has greater area and, therefore, better average performance.

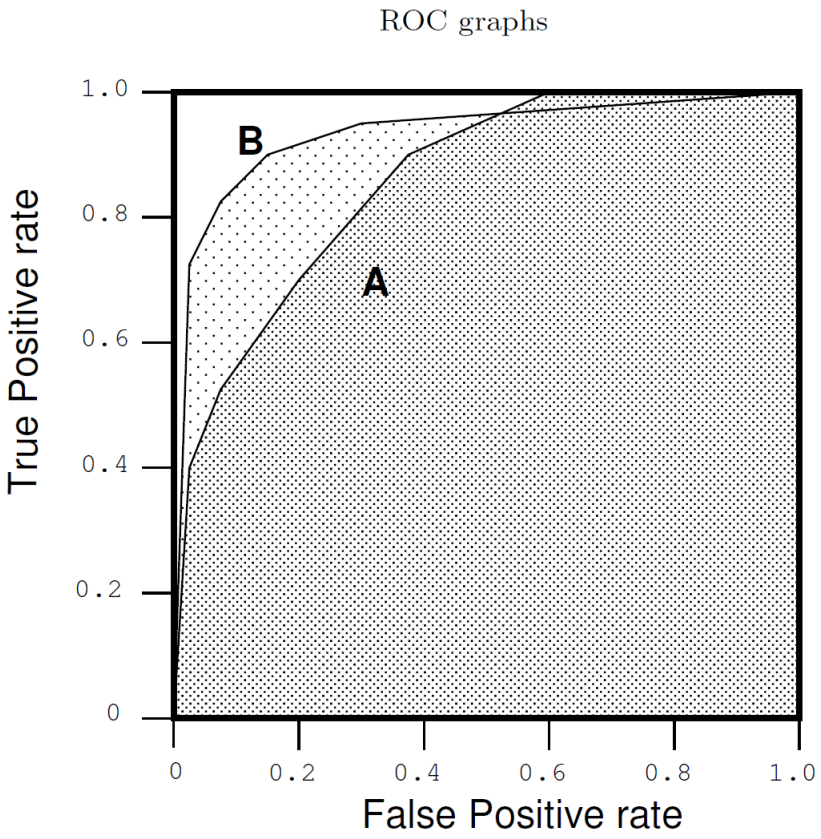

Figure 2.5: An ROC graph that shows the area under two ROC curves for two classifiers A and B.

2.4 Hidden Markov Model (HMM)

A HMM is a model in which the system being modeled is assumed to be a Markov process with unknown parameters. HMM builds a causal model for observation sequence $O = (o_1, o_2, , o_n)$ by introducing corresponding 'hidden states' $q = (q_1, q_2, , q_m)$. The parameters of the HMM are $\lambda = (a, b, \pi)$ where $'a'$ is the parameter for the transition model $P(q_t|q_{t-1})$ and $'b'$ is the parameter for the observation model $P(o_t|q_t)$. The hidden parameters are determined from the observable parameters. The extracted model parameters can then be used to perform further

analysis; for example, pattern analysis or feature subset selection.

For better understanding the HMM, let us consider an example of *persons and stick model*. Assume there are N persons in a closed room (see Figure 2.6). Each person is holding a number of colored sticks. The sticks are of M distinct color. Now, each person throws the colored sticks, one after another, out of the room. At this stage, the only visible outcome is the sequence (order) of colored sticks received. We do not know who, out of N persons,) has thrown which colored sticks.

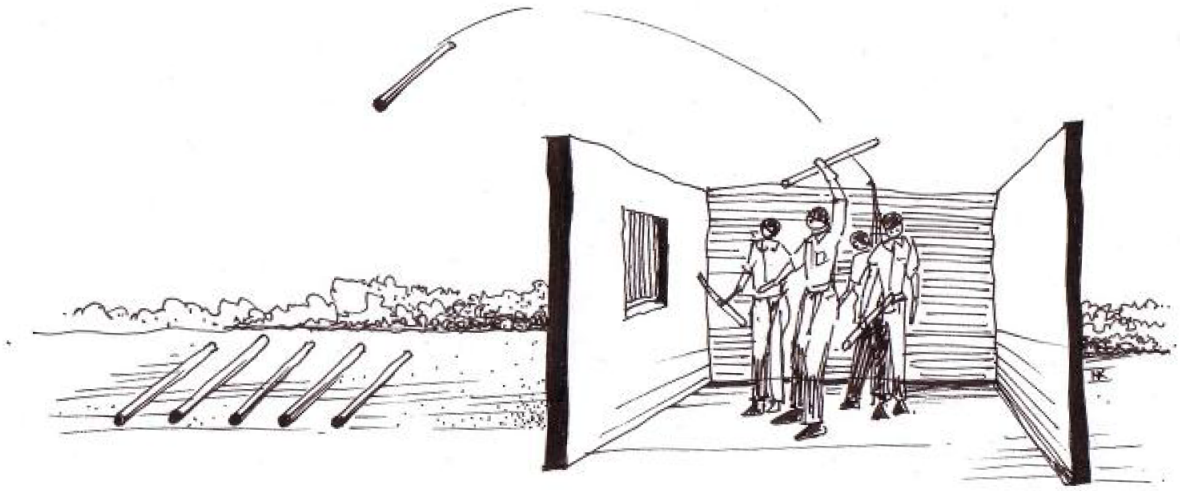

Figure 2.6: The person and stick model

In this example, both the selection of a person and color are completely random. The individuals are hidden from the view, making it a hidden process. The overall process generates a sequence of colors/colored sticks, which forms the output observation sequence. For this process, we do not know the sequence in which a person is throwing the colored sticks, nor the sequence of persons throwing the sticks. The output sequence is very much dependent on the transition probabili-

ties between the various persons/states, and the choice of initial person/state at the beginning. This example can be associated with a HMM, where the states are hidden (like the persons in the example) and the output is the sequence of observations.

2.5 Baum-Welch expectation maximization algorithm

The BaumWelch algorithm [(22)] is a specific case of a generalized expectation-maximization algorithm that can compute maximum likelihood estimates and posterior mode estimates for the parameters (transition and emission probabilities) of a HMM, when given only emissions as training data.

It works as follows: For a given cell S_i in the transition matrix, all paths to that cell are summed. There is a link (transition from that cell to a cell S_j). The joint probability of S_i , the link, S_j and can be calculated and normalized by the probability of the entire string. Let us denote this as χ.

Then we calculate the probability of all paths with all links emanating from S_i and normalize this by the probability of the entire string. Let us denote this as σ.

Finally we divide χ byσ. This is dividing the expected transition from the cell S_i to the cell S_j by the expected transitions from S_i . As the corpus grows, and particular transitions are reinforced, they will increase in value, reaching a local maximum.

Chapter 3

Existing Techniques

In the diagnosis of cancers, gene selection is widely used to select target genes. Due to the high dimensions and relatively small number of samples of microarray data, over-fitting is a common problem. One of the primary goals of gene selection is to avoid this over-fitting problem [25]. In cancer classification, only informative genes which are highly related to particular classes (or subtypes) should be selected theoretically [26].

The primary challenge in microarray data analysis is to select informative genes that clearly differentiate the given classes. In each experiment, since the number of informative genes is very small compared to the total number of genes, utilizing a better search technique is critical. We have divided such approaches into two main categories: statistical approaches and artificial intelligence approaches. *Lazar et al.* [27] and *Saeyss et al.* [28] presented a detailed discussion on filtering techniques for feature selection in bioinformatics.

This chapter is organized as follows. Section 3.1 discusses the recent techniques

that identified important genes using the statistical methods. In Section 3.2, we describe the recent and popular techniques for gene selection using artificial intelligence (AI) techniques.

3.1 Statistical approaches

The methods in the statistic approach rank (score) the discriminability of each gene based on its own gene expression patterns. For the estimations of discriminability, both parametric and nonparametric approaches for have been proposed.

The parametric estimation approaches evaluate the discriminability of genes using a variety of statistical analysis, including signal-to-noise ratio (SNR), t-Test, and least significant difference (LSD). Parametric estimation depends on exact expression levels and the number of replicate samples. The statistical criteria are based on the assumption that the data comes from some kind of distribution [25]. Each parametric approach puts different weights on the variance and number of samples of the criteria. A gene is considered more informative if it possesses a larger corresponding score [25].

3.1.1 Parametric approaches

Signal-to-Noise Ratio (SNR)

Each dataset comprises m samples and n genes. The gene expression data is normalized by subtracting the mean (signal) and then divided by the standard deviation of the expression value (noise) for each gene g_i. Each sample is labeled as (+1, -1) for classification e.g. normal or cancer. The following formula is used

to calculate each gene's 'F' *score*.

$$F(g_i) = \frac{|\mu^{+1}(g_i) - \mu^{-1}(g_i)|}{\sigma^{+1}(g_i) + \sigma^{-1}(g_i)}$$

where μ and s denote the 'mean' and the 'standard deviation' respectively of samples in each class (+1, -1) individually. These genes can then be ranked and top n genes can be selected as the features.

t-Test

The t-Test evaluates whether the means of two groups are statistically different from each other [25]. Since samples may be derived from different experiments and may have different distributions in microarray data analysis, the unpaired two-sample t-Test is often used. The discriminative power of the i^th gene using t-Test is calculated as

$$T(g_i) = \frac{|\mu^{+1}(g_i) - \mu^{-1}(g_i)|}{\sqrt{\frac{\sigma^{+1}(g_i)^2}{M^{+1}-1} + \frac{\sigma^{-1}(g_i)^2}{M^{-1}-1}}}$$

where $M+$ and M- are the sample sizes and μ and s are the respective mean and standard deviation of samples in each class (+1, -1). These genes can be ranked with a T score. Finally the top n gene sets can be selected as the features.

Least Significant Difference (LSD)

Also known as the Fisher criterion, LSD is a classical measure to evaluate the degree of separation between two classes. It is a t-Test-like statistic. The score for gene i is calculated as

$$F(g_i) = \frac{|\mu^{+1}(g_i) - \mu^{-1}(g_i)|}{\sigma^{+1}(g_i)^2 + \sigma^{-1}(g_i)^2}$$

3.1.2 Non-parametric approaches

In contrast to the parametric approach, nonparametric approaches rank samples of each gene using their expression level. These approaches punish the disorders that tend to damage a perfect sample split. The smaller the score a gene receives, the lesser the punishment. This means that a gene is more informative if it has a smaller corresponding score.

Threshold Number of Misclassification (TNoM)

The basic assumption in TNoM is that an informative gene has different values between the two classes. Therefore, these genes are separated using a threshold value. In order to score the given gene and predict the unknown class, a decision rule corresponding to a given expression level is used. TNoM looks to select the values in order to minimize the number of errors:

$$Err(a, b|i) = \sum_k \lambda \{\lambda(i) \neq sign(a \times G(i)_k + b)\}$$

$$TNoM(i) = min_{a,b} Err(a, b|i)$$

Then the genes can be ranked based on TNoM score and top n genes can be selected as the features.

Minimum Distance to Modal Ranking (MDMR)

It first ranks all the sample values of a gene and then computes the minimum distance between these ranks and a modal rank. The ranking algorithm is described by [29]. A score is defined as the minimum number of consecutive swaps needed to arrive at a perfect split of two classes. A score of *0* represents the gene that can split two classes exactly. The MDMR score is defined as

$$MDMR(i) = \sum_{p \in -1} \sum_{q \in +1} h\left(x_p - x_q\right)$$

where *h(x)* is the indicator function

$$h(x) = \begin{cases} 1, x > 0 \\ 0, x \leq 0 \end{cases}$$

Then the genes can be ranked based on MDMR score and top n genes can be selected as the features.

Weighted Punishment on Overlap (WEPO)

The WEPO method was proposed by *Chung et al.* [30] to reduce possible loss of information when using the TNoM or MDMR methods. WEPO introduces the z-score into the rank swapping scheme because genes with identical ordered expression data may not have the same discriminative power. For gene k, the expression levels of samples are first normalized via z-score to eliminate the problem of scaling. The z-score is defined as

$$Z_{ki} = \frac{G(i)_k - \mu_i}{MAD(i)}$$

where μ represents the sample mean and MAD represents the mean absolute deviation of gene k. The punished score of each gene is calculated by estimating the overlapping regions of the two classes. The punishment is defined as

$$WEPO(i) = \sum_{p \in +1} \sum_{q \in -1} \psi(z_{pi} - z_{qi}), \psi(z) = \begin{cases} |z|, z > 0 \\ 0, z \leq 0 \end{cases}$$

Principal Component Analysis (PCA)

PCA is a popular method used in multivariate data analysis to reduce the dimensionality of the data and summarize the data in a parsimonious manner. It has become a useful tool in the analysis of mircroarray data.

Assume we have the gene expression data of p genes for n samples. The goal of PCA is to transform the p gene expression values to corresponding principal comopnent scores. Since the rank of $n \times p$ matrix cannot exceed the minimum of n and p, we expect that the top M principal component scores will convey most of the sample information stored in the original p values.

Lets assume that the essential dimensionality of a microarray dataset has been determined as M. Now we need to find a method that select q genes that can preserve the original data structure. The goal is to determine a gene subset whose first M principal compnent scores have the shortest square distance, denoted as $DIST^2$, to the first M principal component scores of the original microarray

dataset.

Let X and $\hat{X}$ represent the orginal dataset and the selected gene subset. Let Y and $\tilde{Z}$ denote the $(n \times M)$ data matrix of top M principal component scores and the M-dimensional approximation of q genes to the original data configuration resepectively. The gene selection procedure is specified below:

1 Initially set $q = p$ and compute the matrix of principal component scores Y. Set $Z = Y$.

2 Obtain and store the principal component scores matrix by iteratively removing each gene from Z.

3 Compute the square distance $DIST^2$ for each matrix and identify the gene x_u that yields the smallest $DIST^2$. Let us denote the corresponding matrix of scores as $\tilde{Z}_u$.

4 Remove gene x_u. Set $Z = \tilde{Z}$ and goto step 2 with $p - 1$ genes. Repeat until only q genes are left.

The above procedure will select the most informative genes from the gene expression dataset.

3.2 AI approaches

Besides the above mentioned statistical approaches, researchers have used artificial intelligence approaches to select high impact genes from cancer datasets. These approaches employ artificial intelligence techniques like neural networks,

genetic algorithms, decision trees, random forests, support vector machines and self organizing maps. In the subsequent sections, we will briefly describe how these techniques are implemented.

3.2.1 Self-organizing maps (SOMs)

Tamayo et al. [31] stated that SOMs have a number of features that make them particularly well suited to clustering and analysis of gene expression patterns. Since they allow one to impose partial structure on the clusters, they are ideally suited to exploratory data analysis [31]. This is in contrast to the hierarchical clustering's rigid structure and the Bayesian clustering's strong prior hypotheses [31]. They also facilitate easy visualization and interpretation. The benefits of using SOMs are that they have good computational properties, are easy to implement, reasonably fast, and scalable to large data sets.

SOMs are constructed as follows. A geometry of "nodes" is chosen - for example, a 3×2 grid. Starting randomly, then the nodes are mapped into k-dimensional space that are then adjusted iteratively. In each iteration, a random data point P is selected and the nodes are moved in it's direction. The node closest to P, denoted as N_p, is moved the most, while other nodes are moved by smaller amounts depending on their distance from N_p in the initial geometry. In this manner, neighboring points in the initial geometry tend to be mapped to nearby points in k-dimensional space. The process is repeated for m iterations.

The mapping of nodes, adjusted by moving points toward P, is calculated by the formula:

$$f_{i+1}(N) = f_i(N) + \tau(d(N, N_p), i)(P - f_i(N))$$

The learning rate τ decreases with distance of node N from N_p and with iteration number i. The point P that is used at each iteration is determined by random ordering of the n data points generated once and recycled as needed. The function τ is defined by

$$\tau(x, i) = \frac{0.02T}{T + 100i} \qquad for \qquad x = \rho(i) \quad and \quad \tau(x, i) = 0$$

Otherwise, T is the maximum number of iterations if radius $\rho(i)$ decreases linearly with i.

3.2.2 Genetic Algorithms (GAs)

Based on the analogy with biology, GAs are adaptive search techniques, in which a set of possible solutions evolves via natural selection [32]. Genetic algorithms have been a popular choice of researchers for feature selection among high dimensional datasets [32; 33; 34].

In genetic algorithms, a solution is represented by a finite sequence of 0's and l's. This sequence is called a "chromosome". In the application of GAs for feature selection, each chromosome represents a subset of features i.e., the k^{th} bit denotes the presence or absence of the k^{th} feature in the dataset. The algorithm manipulates a finite set of chromosomes, called the population, in a fashion resembling the mechanism of natural evolution. In this mechanism, the chromosomes are allowed

to crossover (mate), and to mutate (change). The mating of two chromosomes produces a pair of offspring chromosomes which are offsprings of their parents. A mutation of a chromosome produces a near identical copy with some components of the chromosome altered due to the mutation.

The optimization process is executed in cycles known as "generations". A set of new chromosomes (bit strings) a_i is created through crossover, which is then mutated and finally evaluated during each generation. Only a predefined number of the best (strongest) chromosomes survive to the next cycle of reproduction due to finite population size.

The population is capable of fast adaptations, despite its limited size, which results in rapid optimization of the criterion function (score).

A high-level algorithmic description of the basic method was outlined by *Siedlecki et al.* [34]. Though many variations of this basic method exist, their description captures the primary characteristics of all GAs.

In the above algorithm f is the so-called fitness function and is calculated as

$$\overline{f} = \sum_{i=1}^{n} f(a_i)/n$$

Similar to the biological function of crossover, the genetic algorithms, that were designed for modeling biological evolution, the crossover operator, *crossover(a,b)*, implements exchanges of information among chromosomes a and b. In particular, if a chromosome is represented by a binary string (as in feature selection), crossover can be implemented by randomly choosing a point, called the crossover point, at

Algorithm 1 : Basic Genetic Algorithm pseudo code

1: Construct an initial population set $\prod = \{a_i\}_{i=1,\ldots,n}$
2: **for** $i \leftarrow 1$ to *Number_of_generations* **do**
3: Initialize mating set $M \leftarrow \emptyset$ and offspring O
4: **for** $j \leftarrow 1$ to n **do**
5: Add f$(a_i)/\overline{f}$ copies of a_i to M
6: **end for**
7: **for** $j \leftarrow 1$ to $n/2$ **do**
8: Select a pair a_j and a_k from M and do $O = O \cup$ *crossover(a_j, a_k)* with probability P_c
9: **end for**
10: **for** $i \leftarrow 1$ to n **do**
11: **for** $j \leftarrow 1$ to d **do**
12: Switch the j^{th} bit in $a_i \in O$ with probability P_m
13: **end for**
14: **end for**
15: Update the population $\prod \leftarrow$ *Combine($\prod$, O)*
16: **end for**

which two chromosomes exchange their parts to create two new chromosomes [34]. For instance, given two strings, $0010 - 0101$ and $1011 - 1010$, the crossover operator cut them in the middle. The result, *crossover(00100101, 10111010)* will produce two new chromosomes i.e., 0010-1010 and 1011-0101.

Mutations increases the variability of the population. In the above example, involving bit strings, a mutant can be created by changing at random one or more bits in the structure i.e., 0010-1010 can be mutated as 0010-1011 (last bit inverted) or 1010-1010 (first bit inverted) or any other bit(s) can be inverted to form a mutated copy (the hyphens are used in the notation for ease of understanding).

The new population if formed by combining the old population and the offspring(s), the method is symbolically denoted as $H \leftarrow Combine(H, O)$. There are a number of possible implementations of this procedure, ranging from more radical, like $H \leftarrow O$ i.e., the new population comprises the offspring(s) only, to

less restrictive, like "select n best chromosomes from H and O" which is mostly used.

The population size, crossover rate and mutation rate are common for all implementations. The crossover rate is the probability of accepting an eligible pair of chromosomes for crossover. The mutation rate is the probability of switching bits in the chromosomes. The crossover rate usually assumes high values, close or equal to one, while the mutation rate is typically small (1 to 15%) [34].

The *fitness* function is another key element in the efficient application-oriented version of the genetic algorithm. However, the execution and performance of genetic search is also determined by a number of parameters, some of them specific for distinct implementations of genetic algorithms.

3.2.3 Random Forest Gene Selection (RFGS)

Random forest is a classification algorithm that was developed by *Leo Breima* [35] and uses a combination of classification trees. Each of the classification tree is built using a bootstrap sample of the data, and at each split the candidate set of variables is comprised of a random subset. Thus, random forest uses both bagging and random variable selection for tree building.

The pseudo code for this approach is outlined in Algorithm 2 where X is the cancer's gene expression data (containing S samples G and genes) and the Y^S is the label of each sample. The output of the algorithm is a list of top Z genes.

Algorithm 2 : The pseudo sode of the Random Forest Gene Selection method

Input: $X = \{ x_G^S, Y^S \}$, $S = 1, \ldots, s,$, $G = 1, \ldots, g,$, $Y^S = \{$ -1, 1 $\}$
Output: n top genes

```
Begin
for I ← 1 to S do
    do normalize X
end for
end
for I ← 1 to N do
    while (All genes assigned completely) do
        Randomly assign all genes into M groups
    end while
    for J ← 1 to M do
        Build up a decision tree on each group
        Mark the root of each group
    end for
    end
end for
end
Rank gene following the number of marks for every gene
Select the top Z genes from the ranking list
Confirm the genes with biological evidence from public resources
Calculate the average biological genes found in the top Z genes
```

3.2.4 Support Vector Machines (SVMs)

SVMs were introduced by *Vapnik et al.* [36] and successively extended by a number of other researchers. Their remarkable robust performance with sparse and noisy data makes them a good choice in a number of applications from text categorization to protein function prediction [37].

SVM is a learning machine based on statistical learning theory [20]. When used in classification, SVM separates binary labeled training data by constructing a hyperplane, which separates class members from non-members. The training set consists of two sets of data; one contains data known to be in a certain class, i.e. genes that have a common function, whereas the other contains data that does not. The genes that are in the class are labeled positively and those which are not in the class are labeled negatively.

From this data, SVM can learn to distinguish between expressions of genes in the training set that are in the class and those of genes that are not. They can then classify new genes as being members or non-members using the information already learnt about which gene expressions correspond to which class. A hyperplane that has maximal distance from members to non-members is called a maximum margin hyperplane. When the data is not linearly separable, SVM maps the data into a higher dimensional space,called a feature space, and defines a separating hyperplane. The kernels of the SVMs automatically realize a non-linear mapping to a feature space. The hyperplane found by an SVM in the feature space corresponds to a decision boundary in the input space.

Let x_i, where i=1,2,. . .,M, be feature vectors of a training set X that belong either of the to two classes ω_1 or ω_2. Using this training data, SVM finds an optimal hyper-plane with the maximum margin that can separate the unknown input patterns into the 2 classes. Since many hyperplanes separating the feature vectors are possible, SVM finds the one that has maximum margin and better generalization performance for classification.

$$g(x) = \mathrm{w}^T \mathrm{x} + w_0 = 0$$

Since SVM is basically a linear classifier that classify linearly separable data, there may be scenarios where the feature vectors might not be linearly separable. To overcome this issue, kernel trick is used.

The original input space is mapped into a high-dimensional feature space using kernel functions where it becomes linearly separable. The performance of an SVM classifier is dependent on the choice of a proper kernel function. Different kernel functions have been employed for different classification tasks. Some of the popular kernel functions and their mathematical forumlae are mentioned below.

Polynomial kernel with degree d

$$K(\mathrm{x}_i \ , \ x_j) = (\gamma \ x_i^T \ x_j + 1 \ )^d \ , \ \gamma \ ¿ \ 0$$

Radial basis function kernel

$$K(\mathrm{x}_i \ , \ x_j) = \exp \ (\text{- } \gamma \ \| \ x_i \text{ - } x_j \ \|^2), \gamma \ \text{¿ } 0$$

Sigmoid kernel function (also called known as *Multi Layer Perception Kernel* or *Hyperbolic Tangent Kernel*)

$$K(\mathrm{x}_i \ , \ x_j) = \tanh \ (\ \gamma \ x_i^T \ x_j + \mathrm{r} \ )$$

Gaussian kernel function

$$K(\mathrm{x}_i \ , \ x_j) = \| \ x^i \text{ - } x^j \ \| \ \sigma \ ^2$$

In above formulae, γ, σ, r and d are adjustable kernel functions which are adjusted based on the data

3.2.5 Artificial Neural Networks (ANNs)

Neural networks are made up of interconnected artificial neurons that mimic human brain processing. The interconnection links between neurons carry certain weights. Each neuron's output is determined by using an activation function such as sigmoid and step. In case of neural networks that are trained with training patter of known classes, they are called supervised learning NN.

The supervised learning process of the neural network consists of a unique input signal and corresponding desired output signal. The network is trained

until it reaches a stable state where the synaptic weights doesnt change and maps to their corresponding output.

The encoding of neural network is composed of two parts; a binary string and a weight matrix. The binary string is used to encode the input features, where the value of each bit is 1 or 0 and it's length is equal to the maximum number of input features. Each bit in the string corresponds to a feature i.e., '1' indicates presence of a feature, while '0' indicates absence of the feature.

Since an input feature corresponds to an input node in neural network, the binary string also represents whether the corresponding input nodes exists. The connection weights of the neural network are specified through the weight matrix that has adimension of $(m + h + n)(m + h + n)$. The connection weight from the j^th node to the i^th node is denoted by w_ij. Since there is no connection from the higher-numbered node to the lower-numbered node, only half of the matrix is required. w_{ii} denotes the bias of the i^th node.

The evaluation of fitness is defined by

$$\text{fitness} = 1 - E - P$$

where E is the error rate of classification, P is the penalty factor for the number of a feature subset

The error rate of classification is given by

$$E = 1_{\overline{T.n}} \sum_{t=1}^{T} \sum_{i=1}^{n} (Y_i(t) - Z_i(t))^2$$

where T is the number of training samples, n is the number of output nodes, *Yi(t)* and *Zi(t)* are actual and desired outputs of node i for sample t.

The penalty factor for the number of a feature subset is defined by

$$\mathrm{P} = u \ . \ \mathrm{M}$$

where M is the number of selected features, u is a penalty coefficient which is a small user-defined parameter

Fitness calculates the individual selection for the next generation. First, the n parent individuals and the n child individuals are sorted together according the fitness. Then, the best n individuals are selected from them.

Chapter 4

Literature Review

Through selective reference to some of the existing literature, this chapter provides a clearer understanding of different methods that have been used in the identification of important genes from high dimensional datasets.

Feature selection in high dimensional datasets is the application area in this thesis. Therefore, an extensive review of the literature is performed focusing on those researches that addresses our application area. We present a review of some of the key papers that have been highly influential to the application area. Most of the papers attempted to solve the reported limitations of these techniques while focusing on feature selection. Therefore, after describing the findings of each of the papers reviewed, we report the limitations of each of the respective techniques.

This chapter is organized as follows. Some of the basic statistical approaches are discussed in Section 4.1. In Section 4.2, we discuss some advances statistical approaches employed for gene selection in high dimensional datasets. Section 4.3 enlists some of the biological approaches in the said area. In Section 4.4,

some of the approaches that employed artificial intelligence for gene selection are discussed. A summarized view of all the approaches is presented in Section 4.5

4.1 Basic Statistical Approaches

Among the existing methods that are based on extensive data analysis, the principal component analysis (PCA) [38], t-test [39], Mann-Whitney U test, Chi-squared test are worth mentioning. PCA is efficient if the distinctive genes are linearly related. To enable the PCA to deal with a non-linear relationship, kernels are deployed before using the PCA. Therefore, PCA is actually used to find the linear relationship among the genes after transforming the data using the kernel techniques. Thus, PCA still remains sensitive to dimensionality and is also complex. Furthermore PCA is directly dependent upon the correlation of each gene. The t-test is a parametric test based on the prior assumption of data distribution. In cases of wrong assumption of the distribution, the t-test looses important genes. Another limitation of t-test is that it is directly dependent on the standard deviation of the expression data. Moreover, both PCA and t-test are sensitive to scaling of the data. Unlike t-test, the Mann-Whitney U-test is a non-parametric test that assesses whether the distributions of two samples of observations come from the same distribution or not [40].

4.2 Advanced Statistical Approaches

Other than the previously mentioned basic statistical methods, there are studies that attempt to better understand the genes and find out their linkage with breast

cancer using advanced statistical methods.

Nguyen et al. [41] reduced the dimensions of gene space using Multivariate Partial Least Squares (MPLS). Then either Polychotomous discrimination (PD) or Quadratic Discriminant analysis (QDA) technique was used to select the genes based on the corresponding gene's mean square error (MSE). One disadvantage of using PD is in the case of quasi-complete separation in the data which is burdensome and usually results in poor classification.

Lee et al. [4] proposed a hierarchical Bayesian model for gene selection by employing hidden variables and using a Bayesian mixture prior to performing the variable selection. By assigning a prior distribution over the dimension (number of significant genes) of the model, the size of the model is controlled. To simulate the parameters from the posteriors, it uses a combination of truncated sampling and Markov Chain Monte Carlo (MCMC) based computation techniques. However, since they used a prior distribution, it might result in an unstable algorithm convergence in case of high-dimensional or highly collinear covariates (related genes). Furthermore the algorithm converge is slow due to the aforementioned reasons [42].

Hedenfalk et al. [1] applied hierarchical clustering to group the similar patterns and found that tumors from individual patients/families did not mix with tumors from BRCA1 and BRCA2 mutation carriers, and remained clustered within the groups. It was shown that gene expression profiling can discover novel classes among BRCAx tumors, and differentiate them from BRCA1 and BRCA2 tumors.

However, the number of selected genes is still quite large (51).

Zhou et al. [43] investigated Bayesian gene selection using the logistic regression model. Three different techniques were used for the identification of posterior distribution of the selected genes, namely, Akaike information criterion (AIC), the Bayesian information criterion (BIC) and the minimum description length (Mdl) principle. These proposed methods were tested on different datasets including hereditary breast cancer dataset from Hedenfalk [1]. The experimental results showed that the proposed methods were effective in finding some genes that were consistent with the contemporary biological knowledge.

4.3 Biological Approaches

There are various studies attempting to identify important genes in a high dimensional dataset using pure biological methods like DNA microarray analysis, gene expression analysis etc. This approach is by far the most expensive and time consuming due to the following reasons. The tests are conducted on the actual blood samples from the breast cancer patients. Hence, the process of identifying patients, getting their consent for sample donation, timely administration of drugs to these patients and collecting the blood samples is a very lengthy process that requires a large amount of time. Moreover, these analysis are performed under the supervision of experienced biologists in biological laboratories. All of these factors make this approach quite expensive and time consuming. Some of the popular biological approaches are discussed below.

Veer et al. [44] used DNA microarray analysis on primary breast tumors of

117 young patients, and applied supervised classification to identify a gene expression signature strongly predictive of a short interval to distant metastases in patient without tumor cells in local lymph nodes at diagnosis. They also established a signature that identities tumors of BRCA1 carriers. This gene expression profile was claimed to outperform all contemporary used clinical parameters in predicting diseases outcome. They also provided a strategy to select patients who would benefit from adjuvant therapy. *Sørlie et al.* [45] found that the prognostic impact of the 231 markers, published by *Veer et al.* [44] on the Norwegian cohort, performed less well (47%) in predicting recurrences within 5 years.

Honrado et al. [46] discussed the identification of a discrete number of genes able to predict whether a breast tumor was associated with a germ-line mutation in BRCA2 or BRCA2 and proposed it could prove helpful. The contemporary criteria for the selection of patients for BRCA1 and BRCA2 analysis was only family history. They argued that this criteria was unspecific due to the fact that around 70% of the studied families were not linked to either of these two genes. However, the identification of the genes somatically altered in BRCA1 and BRCA2 associated tumors opened the opportunity to identify new molecular drug targets specific for these hereditary cases.

Parvin et al. [47] described a novel workflow for the discovery of genes that participate in the breast carcinogenesis process. Genes, whose expression are lightly correlated with BRCA1, BRCA2, and other genes, were identified from public data in the Gene Expression Omnibus (GEO) database [GEO:2012:Online].

Then the genes were tested in the laboratory settings to identify which of these genes regulate the same processes as do BRCA1 and BRCA2. Eight genes were found to impact the homologous recombination among the first nine tested genes.

4.4 Artificial Intelligence Approaches

Computers are powerful machines that can perform complex calculations in mere minutes. Moreover, no working lab or experienced biologists are required to analyze a dataset. Different algorithms can be used to analyze the same dataset yielding different results. These are some of the factors that have made this approach very popular over the last few decades. Some of the popular researches that employed this approach are discussed below.

Kim et al. [48] explored all the pairs, triads, quadruples and quintets of genes and identified their predictive power through strong feature sets algorithm. They assessed the predictive ability of their feature sets by using perceptrons owing to the small amount of data they required for design relative to more general classifiers. 20 genes that appeared most often in the lists of strong performing gene sets in different variable subset sizes, were summarized in separate tables for various classifications.

Qizhong [5] proposed a new method of gene selection and classification by using nonlinear kernel support vector machines (SVM) based on recursive performance elimination (RFE). It used two publicly available datasets; Hedenfalks hereditary breast cancer dataset and Acute lymphoblastic leukemia/Acute myelogenous leukemia (ALL/AML) dataset. A list of twenty strongest genes is computed by

using non-linear kernel SVM-RFE from both the datasets.

Raza et al. [20] used two methods, namely multivariate permutation test (MPT) and significant analysis of microarray (SAM), to select significant genes for feature selection. Then support vector machines (SVM) are applied with polynomial, radial and linear kernels to classify the data. The results showed that all the samples were classified correctly and 100% accuracy rate was achieved among BRCA1- BRCA2 and BRCA1-sporadic. However, the misclassification ratio may be increased if testing and training are done without feature selection due to thousands of genes and fewer samples. This may result in the selection of irrelevant genes that will affect the testing accuracy and will result in low classification accuracy.

4.5 Approaches Summary

A summarized tabular view of the above mentioned approaches and the techniques employed is listed in Table 4.1

Table 4.1: List of approaches, techniques employed and authors for gene selection in high dimensional datasets.

Approach Type	Techniques	Authors	Reference
Advanced Statistical	MPLS + (PDA |QDA)	*Nguyen et al.*	[41]
	Hierarchical Bayesian model	*Lee et al.*	[4]
	Hierarchical Clustering	*Hedenfalk et al.*	[1]
	AIC, BIC, MDL	*Zhou et al.*	[43]
Biological	DNA microarray analysis	*Veer et al.*	[44]
	Gene expression analysis	*Honrado et al.*	[46]
	Laboratory settings	*Parvin et al.*	[47]
AI	Perceptrons	*Kim et al.*	[48]
	SVM + RFE	*Qizhong*	[5]
	MPT + SAM + SVM	*Raza et al.*	[20]

Chapter 5

The Proposed Hybrid Method

In this chapter, we describe our proposed hybrid method that comprises a multi-pass feature selection. In pass one, we rank all genes and determine the highly ranked genes one-at-a-time. In the second pass, we investigate the interrelationship amongst the genes using a Hidden Markov Model (HMM) to find the best subset of genes among the ranked genes. Details of these passes are described in the subsequent sections. Figure 5.1 shows the overall working of the proposed method and the different operations in each pass. In Section 7.2 discusses the selected dataset in detail.

This chapter is organized as follows. In Section 5.1, we discuss the ranking of the genes in the dataset using AUC measure. Section 5.2 describes the gene subset selection from the highly ranked genes using HMM. In Section 5.3, we describe the working of pass 1 and pass 2 with the help of a small geneset example for better understanding.

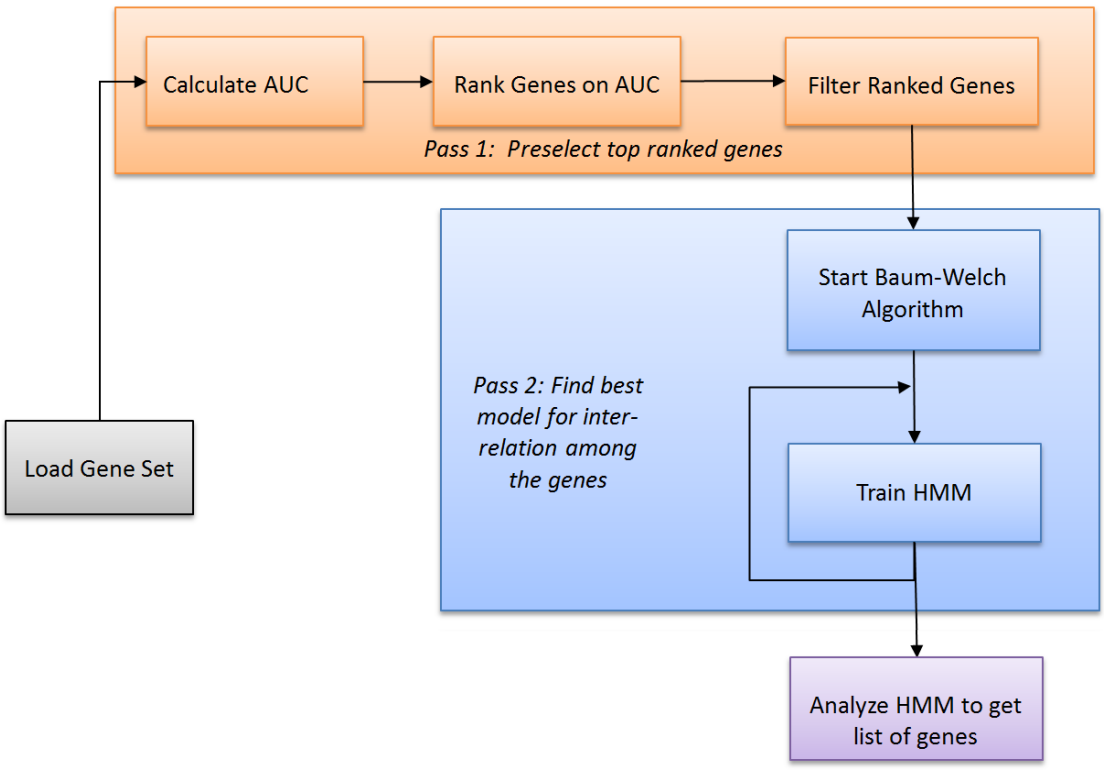

Figure 5.1: The proposed method

5.1 Ranking of genes using AUC

All the genes of the training set were ranked using the first step of *Mamitsuka's* ROC [49]. In a previous study, *Hassan et al.* [40] discussed that this is equivalent of the Mann-Whitney U statistic normalized by the number of possible pairings of positive and negative values. They described that the AUC actually represents the probability that a randomly chosen positive example is correctly rated (ranked) with greater suspicion than a randomly chosen negative example [40].

Let us consider a training gene expression dataset $\boldsymbol{G}$ *of* $\boldsymbol{r}$ *records (patients/samples), where each record comprises* $\boldsymbol{n}$ *gene expressions:* $g_1, g_2, g_3, \ldots, g_n$. *Each of the* $\boldsymbol{n}$ *genes of a record has a differing discriminative power that is re-*

flected by its respective AUC. We calculate this power by plotting the ROC curve for each gene paired with the mutation type i.e., (g_i, M_i)*, where* $1 \geq i \geq n$ *and* $M_i \in \{BRCA1, BRCA2\}$*. We calculate the AUC of the ROC curve and order the genes based on their respective AUCs. The AUC measures discriminative power of genes, i.e., their ability to correctly classify those with and without the disease [50]. Therefore, a high rank of a gene is an indicative of its high likelihood to be associated with BRCA1 or BRCA2 mutation.*

This process begins by calculating the AUC for each gene as described below:

1. Select the gene

2. Sort the gene expression with the class level. This class level is the actual value when calculating the AUC score.

3. Calculate AUC

   (a) For the predicted value, assign +1 incrementally, from the first sample till the last and calculate the AUC for each. (Start by assigning a +1 to the first sample. Then calculate the different measures like TP, FP, Sensitivity, and Specificity for the actual value and the predicted value and calculate the AUC. Next assign +1 to the first 2 samples, calculate the measures and the AUC. Repeat until all samples have +1 as their predicted value, calculate the different measures and calculate the AUC. Finally select the maximum AUC among all the AUCs calculated for various threshold settings)

4. Goto step 1 and select the next gene.

Each gene was ranked according to the scores computed from the ROC curve of the pair of the corresponding gene and the class label as previously described. Comparatively less ranked genes are filtered out by using a threshold level to the AUC scores. For our experiment in this research, we used two threshold values i.e., 0.8 and 0.85 where 0.8 was the least restrictive whereas 0.85 was the most. The best results were achieved by setting the threshold value to $AUC \geq 0.8$. Moreover, we knew that random guessing would yield an AUC value of 0.5, hence the selected threshold value should be greater than 0.5. Furthermore, some previous studies [40; 49] had already used AUC measure to preselect top ranked genes, thus, we got an idea for the minimum threshold from their research as well. Figure 5.2 shows the pictorial view the operations involved in the calculation of calculating the AUC for all the genes.

5.2 Gene subset selection using Hidden Markov Model

This pass begins by receiving a list of important genes that have an $AUC \geq 0.8$ from the previous pass. In our HMM, each state represents a gene. We start training the hidden Markov model and calculate the two-way Wilcoxon Rank-sum measurement [51] for each state (gene) with the other state (gene).

In order to select the best possible model that fits the given data, we use the well-known Baum Welch expectation maximization algorithm [22]. It runs in a loop either for 50 cycles or until the log likelihood value of the new model is less than the log likelihood value of the previous model (indicating no better model

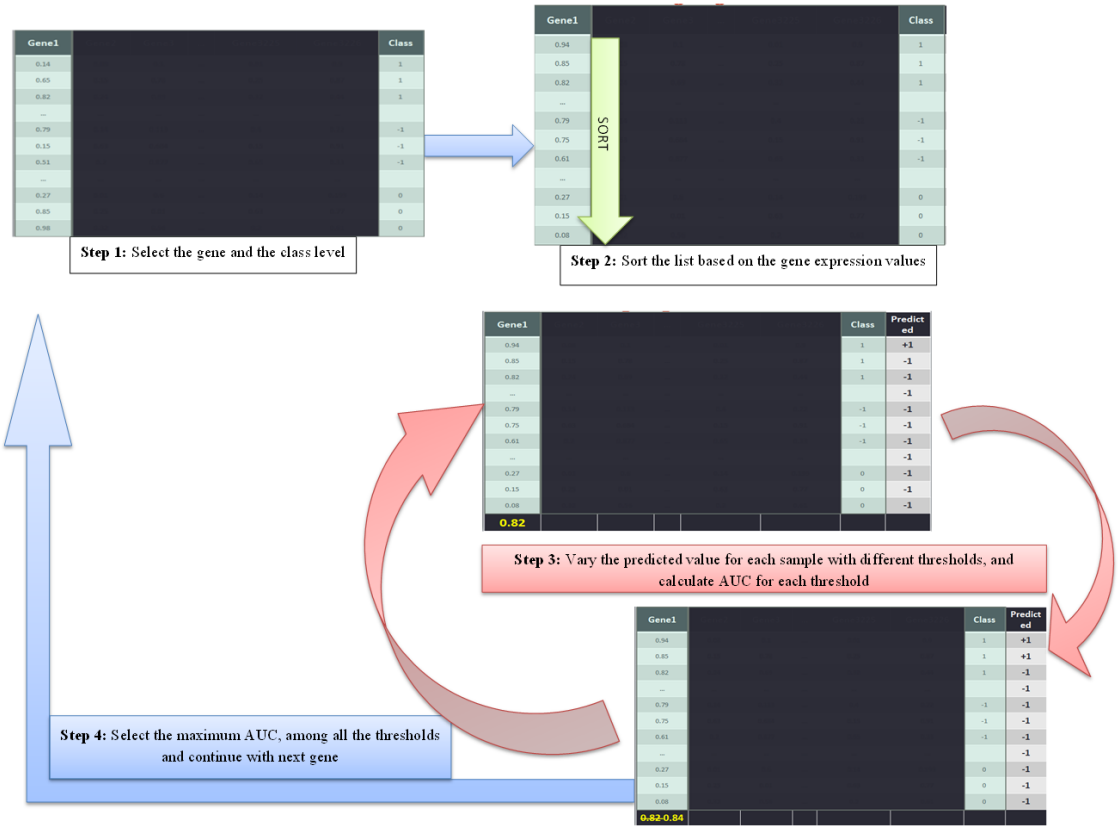

Figure 5.2: Pass 1 - Ranking of genes using AUC

possible). This ensures that the best model is selected among all the possible models that fit the given data. After each cycle, the emission, transition and initial probability (π) matrices are updated so that the next model is better than the current model as shown in Figure 5.3.

While building the model, the p-value for the two-way Wilcoxon ranksum measurement of each gene with all other genes are calculated and saved in a separate matrix referred to as the p-value matrix. Every cell in the matrix is the numerical value of the ranksum measure between two genes' expression vectors (expression profile values). Hence, if we have 20 genes, the matrix will have 20 rows and 20 columns (20x20) where each cell holds the ranksum measure of each

gene with all the other genes. The proposed null hypothesis (H_o) for comparing any two genes is that "two genes are similar if they have a p value greater than 0". Conversely, the alternate hypothesis (H_1) is that "two genes are distinct if the p value is 0".

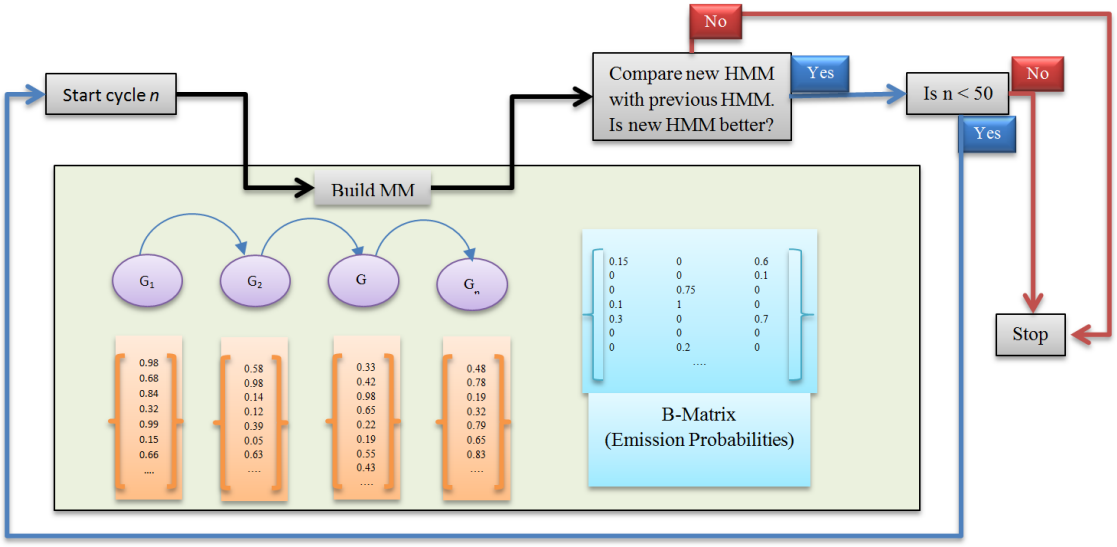

Figure 5.3: Building HMM

Once the best model has been selected, we study the outputs of the hidden Markov model especially the emission probability matrix and the p-value matrix. We identify all rows and columns that have non-zero values and study the pattern in which the values are spread over the matrix. Then we compare the values with the null hypothesis to identify the genes that are distinct as well as those genes that are common. Figure 5.4 shows the p-values matrix and one common and one unique gene. In case of common genes i.e., genes whose p-values are non-zeroes and their patterns are similar in the p-value matrix, we select the gene that has the highest AUC among them.

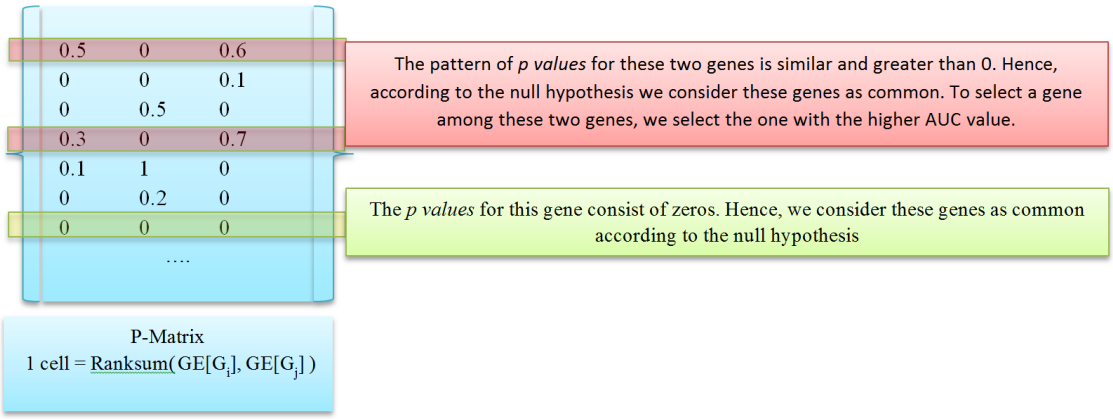

Figure 5.4: Analyzing the $p - value$ matrix

5.3 A small geneset example

For better understanding the various steps in pass 1 and pass 2, let us consider a small geneset of 5 genes each with 10 expression values listed in Table 5.1. Each gene has a class level associated with it; +1 indicate BRCA1 mutation while -1 indicate BRCA2 mutation. Now we input this geneset into the first pass.

Table 5.1: A sample geneset

Gene1	Gene2	Gene3	Gene4	Gene5	Class
0.31	0.63	0.32	0.64	0.91	+1
0.65	0.90	0.14	0.12	0.38	-1
0.88	0.14	0.09	0.97	0.56	-1
0.29	0.93	0.79	0.39	0.48	-1
0.17	0.80	0.84	0.11	0.81	+1
0.23	0.77	0.63	0.83	0.31	-1
0.54	0.10	0.77	0.57	0.17	+1
0.80	0.28	0.91	0.47	0.26	+1
0.14	0.79	0.73	0.17	0.97	+1
0.60	0.18	0.81	0.49	0.16	+1

5.3.1 Ranking of genes using AUC

In this pass, the AUC for all the genes is calculated using the following steps:

Step 1 Select the first gene and sort it according to its expression values with the class label

Step 2 Then assume the predicted values for this gene class by assigning +1 to the first sample and -1 to the rest of the samples. This is threshold 1 i.e., one positive label while the rest are negative labels.

Step 3 Now calculate the different measures like TP, FP, *sensitivity* and $specificity$ for the actual values and the predicted values and then calculate the AUC .

Step 4 Next modify the predicted values of this gene's class by incrementing 1 to the threshold i.e., the first two samples are assigned +1 while the rest are considered -1.

Step 5 Again calculate the different measures and then calculate the AUC for this new threshold. After calculating the new AUC value compare it with the old AUC value for the previous threshold and pick the max AUC.

Step 6 Alter the threshold until all the samples have the same class. Now the AUC value of this gene is the maximum possible value for all the thresholds.

Step 7 Repeat the above mentioned steps for the next gene.

Figure 5.5 explains the first four steps for calculating the AUC for a single gene by varying the predicted class threshold. Steps 5 and 6 ensure that the maximum value of AUC, among all the calculated AUC values for the various thresholds,

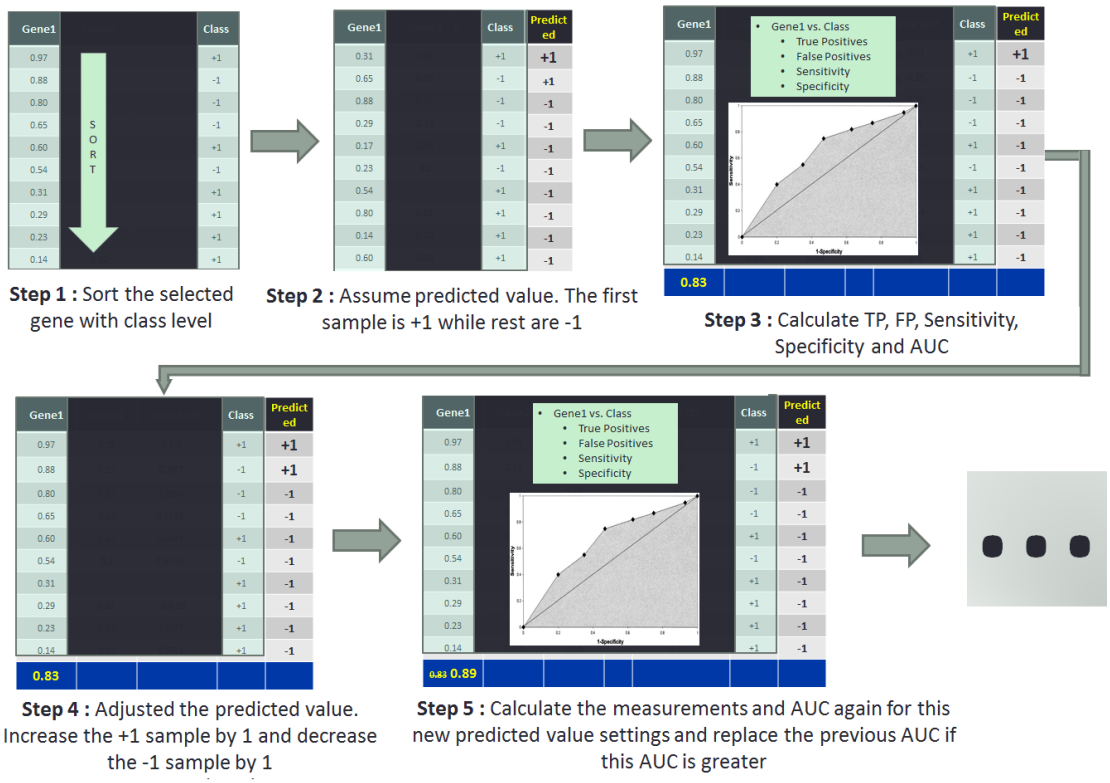

Figure 5.5: Calculating the AUC for a single gene

is selected. Step 7 ensures that the process for calculating the maximum AUC for a gene that has been specified is repeated for all the other genes. After the calculation has been completed for the whole geneset, we will have a dataset in which each gene will have a respective AUC score indicating it's strength. Finally we discard the less important genes by setting a threshold value of $AUC \geq 0.8$. This ensures that only the most powerful genes are selected and transferred for the second pass. Table 5.2 contains the genes that have an AUC value above the threshold and are used in the second pass. The AUC value of each gene is listed in the last row of Table 5.2.

Table 5.2: The filtered geneset and the AUC value for each gene

Gene1	Gene2	Gene3
0.31	0.63	0.32
0.65	0.90	0.14
0.88	0.14	0.09
0.29	0.93	0.79
0.17	0.80	0.84
0.23	0.77	0.63
0.54	0.10	0.77
0.80	0.28	0.91
0.14	0.79	0.73
0.60	0.18	0.81
0.91	0.87	0.80

5.3.2 Gene subset selection using Hidden Markov Model

This pass starts when the list of genes has been filtered in pass 1. The operations in this pass are mentioned below:

Step 1 Initialize the emission, transition and initial probability (π) matrices randomly

Step 2 Start Baum-Welch expectation maximization algorithm for n number of cycles

Step 3 Start training the HMM

Step 4 Once the HMM has been trained, check whether this model is better than the previous model based on log likelihood value

Step 5 If the current model is better, update the matrices that were randomly initialized in step 2. If the current model is not better than the previous

model, then stop

Step 6 Goto step 3 and train a new HMM with the updated matrices

Once the above mentioned process is completed, we get a HMM that best fits the given data. An example of a HMM that fits the geneset example data listed in Table 5.2 is shown in Figure 5.6 where we have 3 genes and the emission probability shown for each gene.

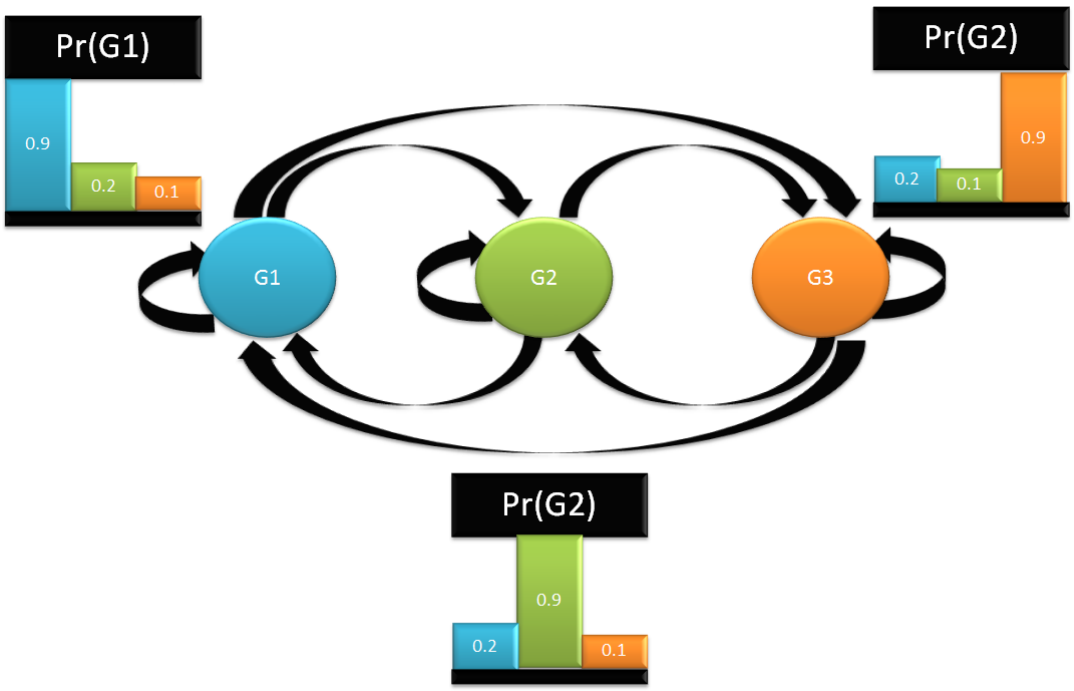

Figure 5.6: A sample HMM for geneset shown in Table 5.2

In order to understand the use of p-values and the other matrices in analyzing the important genes, let us consider a sequence as *'AABABACDDAABBCCDD'*. Figure 5.7(a) shows the Markov process/model (λ) for this sequence using four distinct states, and the corresponding state transition matrix. Figure 5.7(b) shows another HMM (λ^*) that uses five states and the corresponding transition model for the

same sequence. Since, in HMM (λ^*), we have considered the same symbol/feature 'A' twice, the observation emission probability for 'A' from one of the states assigned for 'A' will be zero. Hence, analysis of the transition matrix Z^* reveals that there are at least four unique symbols while the remaining one (out of five states) can be considered as redundant. A similar analysis can be done using the state transition matrix Y^*. By identifying the redundant features amongst the pre-selected features we obtain the most discriminative feature subset.

In the above mentioned example, we have simple characters like 'A', 'B' etc and we can easily identify the recurring elements because each character is a scalar value. However, the pattern (same rows, columns) is helpful in identifying the recurring and/or unique characters.

Now when we try to implement the above mentioned scenario in our gene expression dataset, we have to consider that we have array of genes, in place of single characters, that we can consider as a vector. Therefore, for comparing the two vectors (genes), we have used two-way Wilcoxon Rank-sum measurement [51] of each gene with all the other genes. Similar to the characters in the string example, the patterns (rows, columns) will be helpful in identifying the unique genes.

In conclusion, after the HMM is modeled, we analyze the p-value matrix to identify the list of unique and common genes. As mentioned earlier, the null hypothesis is that the genes with a $p - value > 0$ are common whereas the genes whose p value is 0 is unique. Now to identify the cluster (group) of the common

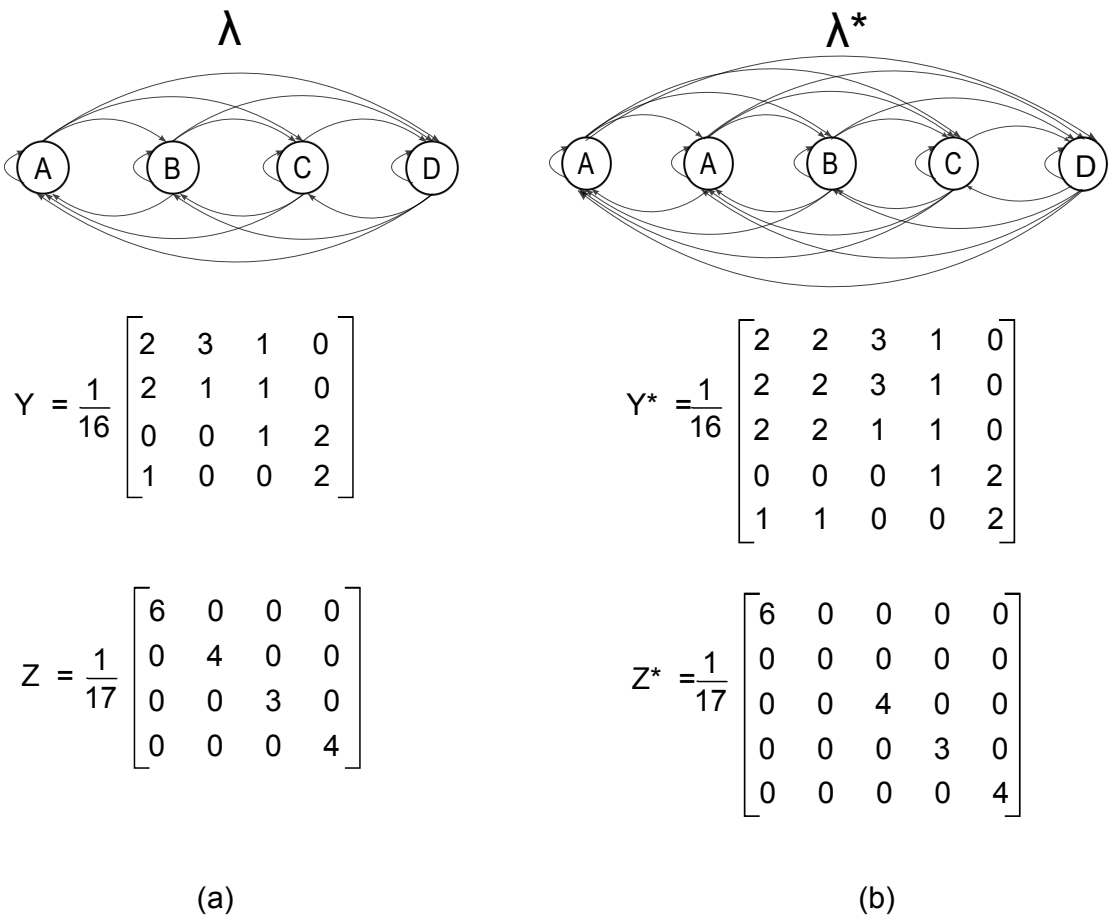

Figure 5.7: (a) A HMM structure for the sequence "AABABACDDAABBCCDD" where each state represents a unique symbol, (b) A HMM structure for the same sequence where an additional state for the symbol is introduced assuming that the symbol 'A' represent two unique symbols. Y (Y*) and Z (Z*) represents the state transition and observation emission probabilities matrices respectively in each case of HMM structure.

genes, we analyze the patterns in the p-value matrix. Figure 5.4 shows the analysis of the matrix for identifying the genes.

Chapter 6

Existing literature results

In this Section, we discuss the results that have been achieved by some of the studies we discussed in Chapter 4. As previously mentioned, a lot of work has been done in the feature selection in dimensional datasets has been done in the last decade. We discuss the results that were achieved by some of these studies. This will help the reader when we compare the list of genes identified by our method with the different list of genes identified by the existing studies.

This chapter is organized as follows. Results from *Hedenfalk et al.* [1] are listed in Section 6.1. The results of *Storey et al.* [2] are discussed in Section 6.2. Section 6.3 has the results from *Zhou et al.* [3]. In Section 6.4, the results from *Lee et al.* [4] are listed. Section 6.5 discusses the results from *Qizhong* [5]. The results of *Xiong et al.* [6] are listed in Section 6.6.

6.1 Results of *Hedenfalk et al.* [1]

Hedenfalk et al. [1] identified *51* genes that can be used to differentiate BRCA1, BRCA2, and sporadic cases of primary breast cancer shown in Table 6.1.

Table 6.1: The list of *51* genes identified as important by *Hedenfalk et al.* [1] with their respective IMAGE clone IDs

IMAGE Clone ID		
897781	51209	42888
139354	949932	38763
809981	784830	366824
841617	26082	840702
823940	46019	564803
29054	247818	137638
810057	214731	73531
950682	236055	32231
26184	197176	274638
344109	566887	234150
36775	725680	701481
341130	823775	838568
417124	293104	31842
429135	46182	666377
44666	307843	50413
340644	366647	345645
246194	212198	810551

6.2 Results of *Storey et al.* [2]

Storey et al. [2] processed the breast cancer dataset and identified a list of important genes. The top *45* genes identified as important are listed in Table 6.2

Table 6.2: The list of top *45* genes identified as important by *Storey et al.* [2] with their respective IMAGE clone ID

IMAGE Clone ID		
566887	814270	838568
212198	46019	365147
42888	26184	75009
784830	344109	295831
366824	236055	136769
950682	548957	812227
51209	810057	133178
32790	841617	841641
897781	898123	49788
247818	31842	197176
82991	366647	23014
809981	236129	291057
29054	711680	134748
293104	293977	127099
949932	840702	36775

6.3 Results of *Zhou et al.* [3]

Zhou et al. [3] identified a list of *20* important genes from breast cancer dataset that are listed in Table 6.3

Table 6.3: The list of *20* genes identified as important by *Zhou et al.* [3] with their respective IMAGE clone IDs

IMAGE Clone ID	
26184	841617
47542	309583
823940	137638
44180	247818
47884	307843
183200	160793
21652	46019
139354	725680
809981	366647
214068	26082

6.4 Results of *Lee et al.* [4]

Lee et al. [4] identified a list of *27* "strongly significant genes for the classification of BRCA1 versus BRCA2 or sporadic" that are listed in Table 6.4 with their respective IMAGE Clone IDs.

Table 6.4: The list of *27* genes identified as important by *Lee et al.* [4] with their respective IMAGE clone IDs

IMAGE Clone ID	
897781	73531
823940	950682
26184	47681
840702	46019
376516	307843
47542	548957
366647	788721
293104	843076
28012	204897
212198	812227
247818	566887
26082	563598
667598	324210
30093	

6.5 Results of *Qizhong* [5]

Qizhong [5] listed *20* genes as important that are shown in Table 6.5.

Table 6.5: The list of *20* genes identified as important by *Qizhong* [5] with their respective IMAGE clone IDs

IMAGE Clone ID	
667598	293104
47884	247818
137638	810899
897781	73531
823940	28012
81331	46019
366824	366647
307843	44180
212198	32790
564803	843076

6.6 Results of *Xiong et al.* [6]

Xiong et al. [6] identified a list of *20* important genes listed in Table 6.6.

Table 6.6: The list of *20* genes identified as important by *Xiong et al.* [6] with their respective IMAGE clone IDs

IMAGE Clone ID	
175123	179804
714106	563444
210887	345423
29054	246194
36775	23014
21652	51209
233721	341130
666377	345645
50413	32790

6.7 Other studies

In addition to the above discussed studies, there are other researches that identified a list of important high-impact genes by *Kim et al.* [48], *Zhou et al.* [43] and *Jazaeri et al.* [52] among others.

Chapter 7

Experiments and Results

This chapter discusses the experimental setup for this research, the dataset that was used and the results that were achieved from our proposed hybrid method. This chapter is organized as follows. Section 7.1 discusses the experimental setup. In section 7.2, we present a brief overview of the selected dataset. Section 7.3 discusses the list of genes that were identified as important by our hybrid method.

7.1 Experimental Setup

The proposed hybrid method was programmed using MATLAB and was executed on a Intel Core i5 machine with 4 GB RAM. The execution time of the program took about 50-60 minutes for writing the output files. These files were then analyzed to identify the important gene set. Each of the gene in the identified geneset was then searched on Database for Annotation, Visualization and Integrated Discovery (DAVID) [7] to identify the list of associated diseases. In addition, several major medical databases like PubMed NCBI [53] and search engines like Google were also used to search for the diseases associated with the identified geneset.

7.2 Dataset

The data used in this research came from the breast cancer cDNA microarray experiment by *Hedenfalk et al.* [1]. Data were collected from the blood of patients with primary breast cancer who had a family history of breast or ovarian cancer, or both, that was compatible with a dominant mode of inheritance. Biopsy specimens of primary breast tumors from patients with germ-line mutations of BRCA1 (seven patients) or BRCA2 (eight tumors from seven patients) were selected for analysis. Gene expression profiles were generated for each sample and were used to determine which of the genes expressed by the tumors correlated with the BRCA1-mutation-positive tumors and the BRCA2-mutation-positive tumors. Lastly 3226 genes were selected that were expressing distinctly between the two tumors.

7.3 Results

Since our dataset was extremely small, we ran the hybrid method twice to find the important genes. In the first run, we did not use any validation technique and got a list of genes that were identified as important. In the second run, we used leave one out cross validation (LOOCV) to validate the small dataset. Since we had only 15 *BRCA1 vs. BRCA2* patients, the LOOCV technique resulted in 15 different lists of genes that were identified as important by the hybrid method. In order to identify only the most important genes, we selected only those genes that were repeated in all or most of the 16 lists (1 without and 15 with LOOCV).

Figure 7.1 shows the list of genes identified without LOOCV, and for each iteration of LOOCV.

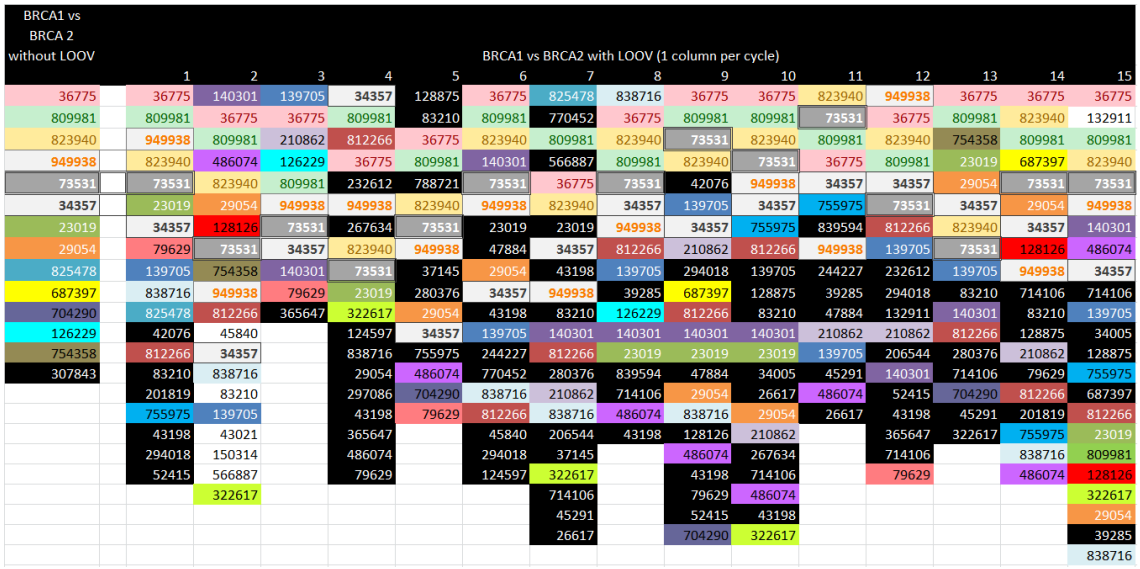

Figure 7.1: List of genes without and with LOOCV

Table 7.1 shows the list of selected genes with their respective IMAGE clone ID, Entrez Gene ID and the AUC value for each gene.

Table 7.1: The list of genes selected using the hybrid of AUC-HMM method with their respective IMAGE clone ID, Entrez Gene ID and the AUC value for each gene

IMAGE Clone ID	Entrez Gene ID	Entrez Gene Symbol	AUC value
36775	3030	HADHA	0.964
823940	10140	TOB1	0.929
949938	1471	CST3	0.929
73531	23479	ISCU	0.929
34357	60	ACTB	0.911
23019	2778	GNAS	0.893
29054	10121	ACTR1A	0.857
809981	2879	GPX4	0.857

Table 7.2 lists the Image Clone ID, Gene Symbol, Gene ID, and the diseases

associated with it according to Database for Annotation, Visualization and Integrated Discovery (DAVID) [7] analysis.

Table 7.2: The list of genes with DAVID's analysis [7]

Image Clone ID	Entrez Gene Symbol	Entrez Gene ID	DAVID's analysis
36775	HADHA	3030	1) HELLP syndrome, maternal of pregnancy 2) LCHAD deficiency
949938	CST3	1471	1) A genome-wide association for kidney function and endocrine-related traits in the NHLBI's Framingham Heart Study. 2) Cerebral amyloid angiopathy 3) Macular degeneration, age-related, 11 4) New loci associated with kidney function
73531	ISCU	23479	Myopathy with lactic acidosis, hereditary
34357	ACTB	60	Dystonia, juvenile-onset
23019	GNAS	2778	1) Acromegaly 2) Genome-wide association study of theta band event-related oscillations identifies serotonin receptor gene HTR7 influencing risk of alcohol dependence. 3) McCune-Albright syndrome 4) Osseous heteroplasia, progressive 5) Pituitary ACTH secreting adenoma, somatic 6) Pituitary ACTH secreting adenoma, somatic 219090 7) Pseudohypoparathyroidism Ia 8) Pseudohypoparathyroidism Ib 9) Pseudohypoparathyroidism Ic 10) Pseudopseudohypoparathyroidism

Chapter 8

Impact of our work

In this chapter, we have compared the results, achieved from our hybrid method, with other cancer studies biologically and statistically in order to evaluate the impact of our work. The biological significance of the identified genes is discussed; by referencing the identified genes in existing studies, through analyzing transcription factors, and through the biological significance of genes through Protein-Protein interaction network. The statistical significance of the identified genes is calculated based on GSEA measurements.

This chapter is organized as follows. Section 8.1 discuss the biological significance of the identified genes through referencing the existing studies (8.1.1), through analyzing Transcription Factors (8.1.2), through Protein-Protein interaction network (8.1.3) and through finding common genes between our identified genes and genes identified by other studies (8.1.4). The statistical significance of our identified genes is discussed in Section 8.2.1.

8.1 Biological significance of the selected genes

8.1.1 Biological significance as referenced by the existing studies

HADHA: Interestingly, HADHA expression has been identified having a strong association with the risk of breast cancer [54]. The DAVID gene expression analysis tool also reveals that HADHA gene is associated with HELLP syndrome. HELLP syndrome comprises of a group of symptoms that occur in pregnant women who have H (hemolysis i.e., breakdown of red blood cells), EL (elevated liver enzymes) and LP (low platelet count). *M̈uller et al.* [55] reported a statistically significant difference in methylation of adenomatous polyposis coli (APC) [56] in the sera of healthy pregnant women and women who later developed severe HELLP syndrome, perhaps offering a possible tool for early detection of this severe disease in pregnancy. They also described for the first time in a phenomenlogic way that methylations changes in the sera of women in early pregnancy are similar to those in the sera of patients with advanced breast cancer. These evidently justify selection of this gene (related to HELLP syndrome, that is different in terms of its expression in between BRCA1 and BRCA2) as it causes advancement of breast cancer.

TOB1: TOB1 is a protein that in human is encoded by the TOB1 gene. *O'Malley et al.* [57] have shown that TOB1 is expressed in the normal breast epithelial cells (the thin tissue forming the outer layer of a body's surface and lining the alimentary canal and other hollow structures) and that the reduction or loss of TOB1 expression is associated with breast cancer progression. *Helms et*

al. [58] showed a significantly shorter distant metastasis-free survival for breast cancer patients with high TOB1 expression levels.

CST3: Cystatin C or cystatin 3 is a protein encoded by the CST3 gene, and is mainly used as a biomarker of kidney function. The cystatin superfamily encompasses proteins that contain multiple cystatin-like sequences. Some of the members are active cysteine protease inhibitors, while others have lost or perhaps never acquired this inhibitory activity. *Yano et al.* [59] reported a significant reduction in relative cystatin C expression to the cathepsin B expression which was detected in breast cancer tissue. The study concluded that cystatin C expression plays a role in cancer invasion and metastasis. In another study *Vigneswaran et al.* [60] showed for a limited set of tumor samples that cystatin M and C are expressed in primary breast cancer with lymph node metastasis. It concluded that increased expression levels of both cystatin M and C are significantly correlated with larger tumor size in breast cancer. Similarly *Tumminello et al.* [61] reported that cystatin C levels were significantly higher in breast cancer patients than in controls or patients with osteoporosis.

ACTB: Beta-actin (gene name ACTB) is one of the six different actin isoforms which have been identified in human. *Ambrosino et al.* [62] identified that the actin network plays a role in nuclear ER-alpha actions in breast cancer cells, including coordinated regulation of target gene activity, spatial and functional reorganization of chromatin, and ribosome biogenesis. ACTB gene has also been identified as one of the significant genes to quantify urokinase plasminogen

activator in breast cancer [63]. It is worth mentioning here that the urokinase plasminogen activator is a well known system that causes the progression of malignant cancer at various steps [64].

ISCU: Iron-sulfur cluster assembly enzyme ISCU, mitochondrial is a protein that in human is encoded by the ISCU gene. *Favaro et al.* [65] found a highly significant inverse relationship of miR-210 to ISCU expression in 216 patients with breast cancer.

GPX4: Glutathione peroxidase 4, also known as GPX4, is an enzyme that in human is encoded by the GPX4 gene. *Udler et al.* [66] reported that the common variation in GPX4 is associated with the prognosis after a diagnosis of breast cancer. In another study, *Heirman et al.* [67] reported that GPX4 overexpression drastically impeded solid tumor growth in certain cells.

ACTR1A: Alpha-centractin is a protein that in human is encoded by the ACTR1A gene. The most abundant subunits of dynactin, which is associated with the transport of p53 to the nucleus, are encoded by ACTR1A. Disruption of this ACTR1A-dynactin complex via mutations in ACTR1A could potentially result in p53 inactivation, which is intriguing, given the absence of known inactivating mutations in p53 in malignant pleural mesothelioma (MPM) tumors [68].

8.1.2 Biological significance through analyzing of transcription factors

Transcription factors are specialized proteins that bind to specific sites on DNA and turn on or off the expression of different sets of genes. The basic structure of every transcriptional factor mainly contains a DNA-binding domain and an

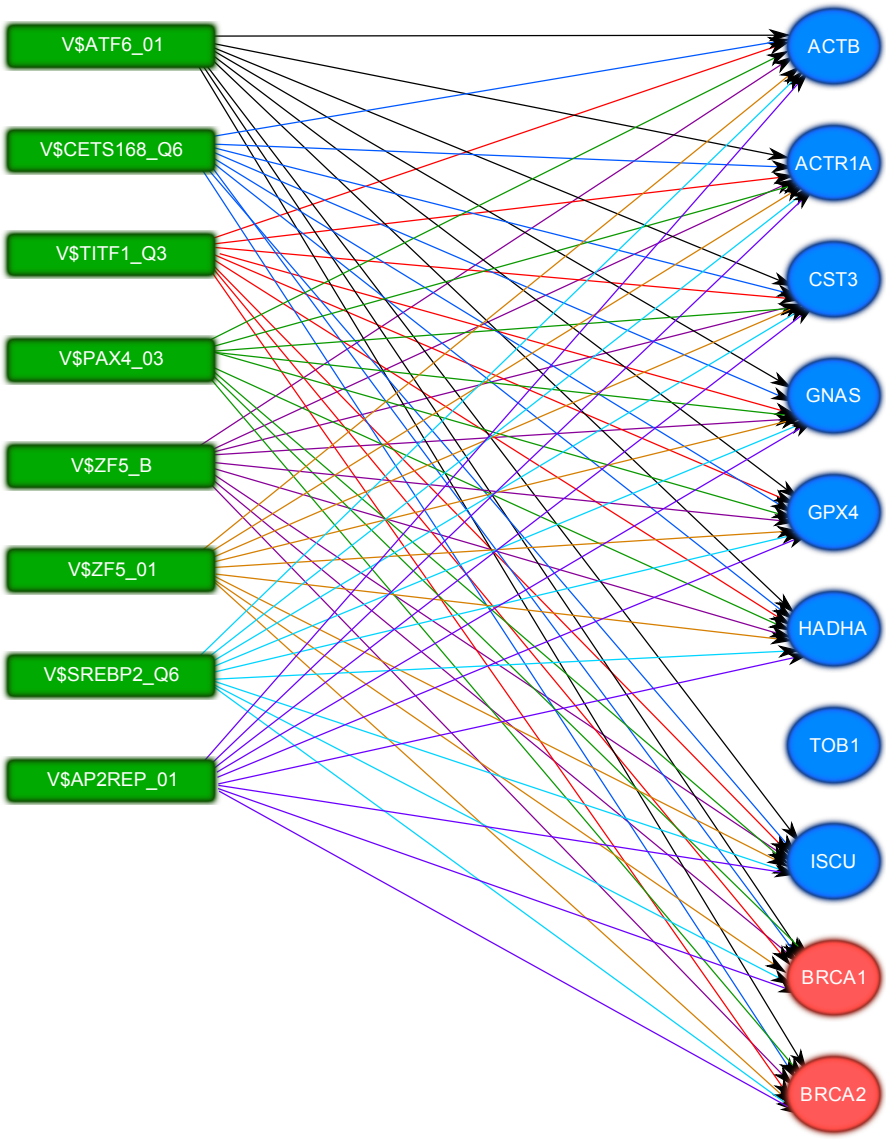

Figure 8.1: Transcription factor network linked with the selected gene set

activator domain [69]. For improvement of disease treatment strategies, recent studies are concentrating on the identification of different transcription factors and blocking them. Transcription factors are known to play a key role in regulating pathways, by controlling the levels of pathway gene expression [70]. We used TRANScription FACtor database's (TRANSFAC) [71] transcription factors that were in GMT (Gene Matrix Transposed) format. In order to find those tran-

scription factors that cover the identified 8 genes and BRCA1 and BRCA2, we used minimum hitting set algorithm [72] i.e. we selected only those transcription factors that covered the minimum number of genes including the identified genes. We found 8 different transcription factors that covered the identified genes and BRCA1 and BRCA2 as shown in Figure 8.4. One of the transcription factors that cover the identified genes, ATF6 (activating transcription factor 6 [73]) has already been linked with cancer [74]. In summary, the transcription factors that cover the identified 8 genes also control the breast cancer genes i.e. BRCA1 and BRCA2.

8.1.3 Significance of the genes based on Protein-Protein interaction network

Figure 8.2 shows the Protein-Protein interaction network for the identified 8 genes, their other neighboring genes and BRCA1 and BRCA2. For simplicity and visibility we have removed the genes that form small clusters, and have renamed the cluster with the corresponding gene symbol that is at the center.

After analyzing the Protein-Protein interaction network, we identified 5 genes, namely ACTB, TOB1, HADHA, GNAS and ACTR1A that form individual clusters and are not directly linked with BRCA2 or BRCA2. Therefore, these genes need to communicate with BRCA1 or BRCA2 indirectly i.e., through other genes. It is interesting to note that 6 genes, namely SMARCC2, POLR2A, H2AFX, SMARCA4, MGMT, SMAD3 and SKP2, provide the communication between BRCA1 and BRCA2 and the other genes. Hence, these genes serve as a "barrier"

for BRCA1 and BRCA2 since they can stop any undesirable communication to BRCA1 and BRCA2.

Figure 8.2 shows that TOB1 is indirectly linked with both BRCA1 and BRCA2 through ESR1, SMAD3 and SKP2. Hence, it is a possibility that TOB1 communicates with either BRCA1 or BRCA2 or both through these 3 genes. If they can be suppressed, the communication between TOB1 and BRCA1 and/or TOB1 and BRCA2 will be broken. Another interesting fact is that TOB1 forms a cluster with 10 other genes that is mostly dominated by SMAD variants (1, 4, 5 and 9), with itself at the center. This leads to an assumption that these 10 genes are very closely correlated with TOB1 and can influence TOB1. Under their influences TOB1 may overexpress/underexpress while communicating with BRCA1 or BRCA2 or both through ESR1, SMAD3 or SKP2.

Figure 8.2 also shows that ACTB is indirectly linked with BRCA1 and BRCA2 through SMARCC2. Hence, it is possible that ACTB communicates with BRCA1 or BRCA2 or both through this gene. Hence, this indirect communication can be broken by silencing/suppressing this mediator gene. In addition, ACTB forms a cluster of 33 genes that is mostly dominated by CCT variants (6a, 7, 3, 4, 5, 8) with itself at the center. If these genes are silenced, the link between them and BRCA1 or BRCA2 or both can be broken. Another interesting fact is that ACTB is directly connected with CST3 and indirectly connected with ACTR1A cluster through SPTBN2. There is a strong possibility that CST3 or ACTR1A can influence ACTB or vice versa.

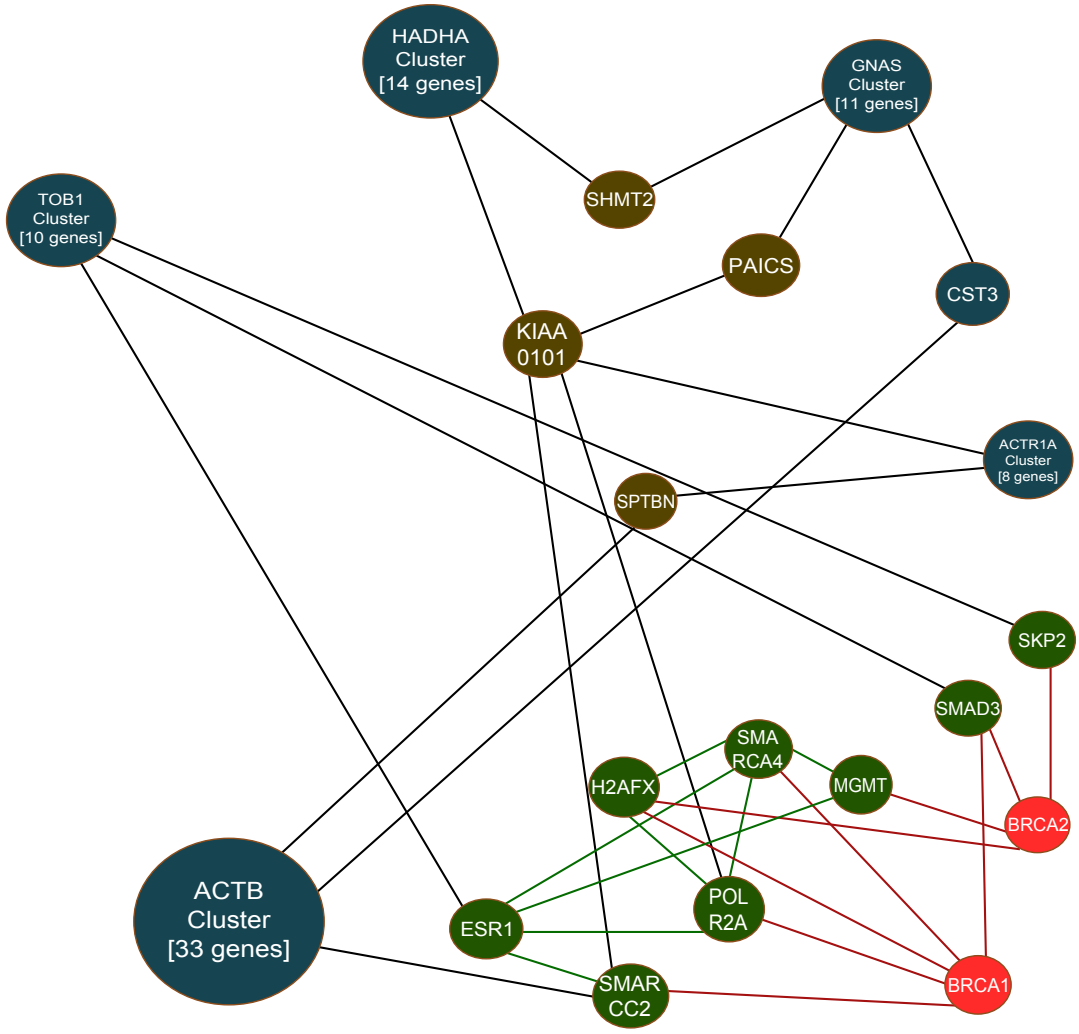

Figure 8.2: Protein-Protein interaction network reflecting the linkage between the selected gene set and BRCA1 or BRCA2 or both

Another identified gene ACTR1A forms a cluster with 8 genes that is mostly dominated by DCTN variant genes (1, 2, 3, 4) with itself and DCTN4 at the center. ACTR1A does not communicate with any gene that is a neighbor of BRCA1 or BRCA2 or both. However, it communicates with ACTB through SPTBN2. If ACTR1A is influencing BRCA1 and BRCA2 or both, this can be stopped by suppressing the SPTBN2 gene. Similarly ACTR1A is directly linked with KIAA0101 that itself is connected with SMARCC2 and POLR2A. Furthermore,

KIAA0101 is also connected with HADHA and GNAS. Hence, if HADHA or GNAS or ACTR1A are being influenced by genes in their respective clusters or by each other, then the link can be broken by silencing KIAA0101.

GNAS forms a cluster of 11 genes with itself at the center. It is directly linked with CST3 which can influence it or vice versa. Moreover, GNAS is indirectly linked with ACTB through CST3. This connection can be broken by silencing CST3 in case of any overexpression/underexpression of either of the three genes. GNAS is also connected with HADHA through SHMT2 and both of these genes can influence each other. This can be stopped by silencing SHMT2. In addition, GNAS is indirectly connected with KIAA0101 through PAICS. If it is influencing KIAA0101, then by silencing PAICS this connection can be broken.

HADHA forms a cluster of 14 genes with itself at the center. HADHA is directly connected with KIAA0101 that itself is connected with GNAS, ACTR1A, POLR2A and SMARCC2. HADHA might be influenced by GNAS or it can influence KIAA0101. This can be stopped by silencing KIAA0101.

8.1.4 The genes that have been already identified by other cancer literature

In order to compare our findings with other cancer literature, we find the common genes among the identified genes using our approach and those identified by other cancer literature. Out of the *8* identified genes using our hybrid method, *5* genes have been mapped with existing literature. Table 8.1 shows a summarized view of the existing studies and the list of genes that are common between them and

our identified list.

Table 8.1: A summarized view of existing studies and their common genes with our identified list

Author(s)	Genes	Common Genes Count	Total Genes	Reference
Hedenfalk et al.	*HADHA, TOB1, ISCU, ACTR1A, GPX4*	*5*	*51*	[1]
Storey et al.	*HADHA, ACTR1A, GPX4*	*3*	-	[2]
Lee et al.	*TOB1, ISCU*	*2*	*27*	[4]
Qizhong	*TOB1, ISCU*	*2*	*20*	[5]
Zhou et al.	*TOB1, GPX4*	*2*	*20*	[3]
Xiong et al.	*HADHA, ACTR1A*	*2*	*20*	[6]

Out of the *8* identified genes by our hybrid method, at least *3* genes ($HADHA$, $ACTR1A$ and $GPX4$) were identified as important by *Hedenfalk et al.* [1] and *Storey et al.* [2]. Moreover, *2* genes ($TOB1$ and $ISCU$) were also identified as important by *Hedenfalk et al.* [1], *Lee et al.* [4] and *Qizhong* [5]. Furthermore, *2* genes ($TOB1$ and $GPX4$) were also listed as important by *Hedenfalk et al.* [1] and *Zhou et al.* [3]. Lastly, *2* genes ($HADHA$ and $ACTR1A$) were listed as important by *Hedenfalk et al.* [1] and *Xiong et al.* [6]. For the purpose of clarity and easy understanding, Figure 8.3 shows a stacked diagram of the genes identified using the hybrid method that were also identified as important by other studies.

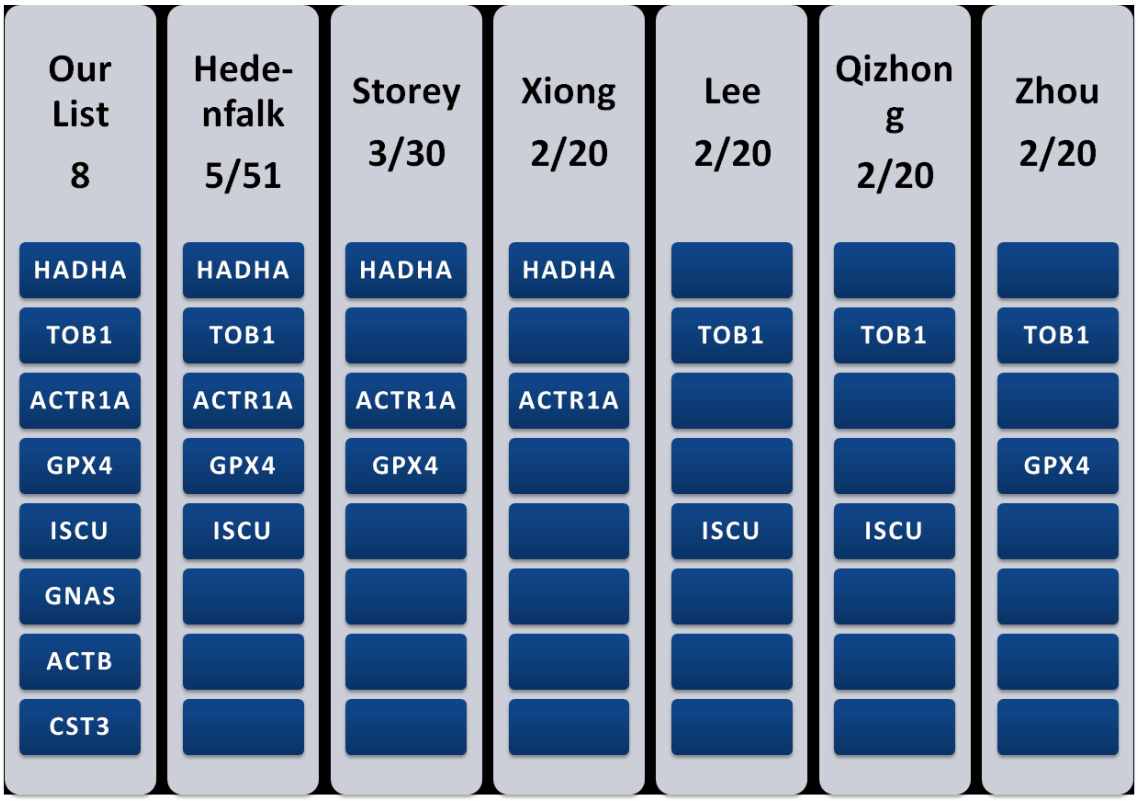

Figure 8.3: A stack diagram (grouped list) showing the genes that are common between genes identified by other cancer studies and our identified gene set and the total genes for each study (common genes/total genes).

8.2 Statistical significance of the selected genes

8.2.1 GSEA measurement

It is interesting to consider the results of gene set enrichment analysis (GSEA) on our set of 8 genes, identified by the hybrid method, against the other gene sets proposed by *Hedenfalk et al.* [1] (51 genes), *Lee et al.* [4] (27 genes), *Qizhong* [5] (20 genes) and 8 other studies. Five out of the 8 identified genes are enriched in BRCA1, while the remaining 3 are enriched in BRCA2. Moreover, the identified gene set has an enrichment score of 0.52 (fourth best in the list and better than *Zhou et al.* [43], *Mamtani et al.* [54] and *Mao et al.* [75]), and members of the

Author	SIZE	ES	NES	LEADING EDGE
Lee et al.	19	0.91	2.19	signal=85%
Zhang et al.	15	0.91	2.12	signal=81%
Hedenfalk et al.	27	0.61	1.95	signal=56%
Mao et al.	16	0.48	1.51	signal=44%
Zhou et al.	14	0.50	1.34	signal=52%
Our list	8	0.52	1.27	signal=72%
Mamtani et al.	10	0.35	0.91	signal=86%
		4th	6th	4th

Figure 8.4: Gene Set Enrichment Analysis (GSEA) for our genes and other studies

leading edge subset (i.e., tags = 63%, list = 14% and signal = 72%) is fourth best and better than *Hedenfalk et al.* [1], *Zhou et al.* [43], *Mamtani et al.* [54] and *Mao et al.* [75]. Furthermore, out of the 8 identified genes, 2 ($ACTR1A$ and $HADHA$) are in the top 50 rank in gene list ($ACTR1A$ is at third and $ISCU$ is at seventeenth position) computed by GSEA with an enrichment score of ≥ 0.95. The most important point to note here is that the 8 genes identified by the hybrid method is the 'smallest' size of gene set in the GSEA analysis. These comparisons indicate that identified gene set contains mostly those genes contributing to the enrichment score, compared to the other gene sets that contain only a fraction of genes contributing to the enrichment score.

Chapter 9

Conclusion

We have proposed a hybrid feature selection method in this paper and used it to identify important genes that are biological significant using their expression levels. We have used the area under the ROC curve (AUC) measure and hidden Markov model (HMM) to obtain a small subset of important genes. Here we have set AUC threshold to ≥ 0.8 to filter all less important genes. The filtered gene list is then forwarded to the HMM that has 103 states. After the HMM is modeled, we analyze the observation emission and state transition probabilities to identify the important gene subset. In order to evaluate the impact of our work, we have compared our results with other cancer studies biologically and statistically. The biological significance of the identified genes is discussed; by referencing the identified genes in existing studies, through analyzing transcription factors, and through the biological significance of genes through Protein-Protein interaction network. The statistical significance of the identified genes is calculated based on GSEA measurements. The results of these comparisons confirm that our method

identified the high impact and biologically relevant genes that contribute towards discriminating BRCA1 mutation from BRCA2 mutation in hereditary breast cancer.

9.1 Significance of our work

We believe our work can help biologists and researchers in the following ways:

- Biologists can perform an in-depth study of identified genes in order to come up with new medicine that can either cure the disease or at the very least restrict the growth of the disease, resulting in an extended life of the patient, in case the cure is not possible.
- Identification of important genes and their rigorous study can result in new methods of detection of the disease so that early prognosis can be done and the affected patient can be informed at the earliest, or the disease can be predicted before it actually affects the patient.
- Researchers can use the algorithm to find important genes in a dataset
  - Without the use of laboratories and expensive equipment,
  - Without waiting for lab results

9.2 Limitations

Although the research has reached its aims, there were some unavoidable limitations. First, the sample size in the dataset was extremely small i.e., 22 patients out of which only 15 were related to over problem. Secondly, although we used

Leave one out cross-validation (LOOCV) to maximize the available samples in the dataset and avoid any overfitting, still the sample size was extremely small. Hence, the proposed method may suffer from overfitting. Thirdly, the proposed method may be successtible to noise. Fourthly, the comparison done with other techniques might not be very relevant since those techniques identified genes individually whereas our method is identifying genes collectively. Finally, since we do not have access to a wet laboratory, we are unable to genetically justify the selection of the identified genes.

REFERENCES

[1] I. Hedenfalk, D. Duggan, Y. Chen, M. Radmacher, M. Bittner, R. Simon, P. Meltzer, B. Gusterson, M. Esteller, O.-P. Kallioniemi, B. Wilfond, ke Borg, and J. Trent, "Gene expression profiles in hereditary breast cancer," *The New England Journal of Medicine: Genomic Medicine*, vol. 344, no. 8, pp. 539–548, 2001.

[2] J. Storey and R. Tibshirani, "Statistical significance for genomewide studies," *Proceedings of the National Academy of Sciences*, vol. 100, no. 16, pp. 9440–9445, 2003.

[3] X. Zhou, X. Wang, E. Dougherty, X. Zhou, X. Wang, and E. R. Dougherty, "Gene selection using logistic regressions based on aic, bic and mdl criteria," *New Mathematics and Natural Computation*, vol. 1, no. 1, pp. 129–145, 2005.

[4] K. E. Lee, N. Sha, E. R. Dougherty, M. Vannucci, and B. K. Mallick, "Gene select: Bayesian variable selection approach," *Bioinformatics*, vol. 19, no. 1, pp. 90–97, 2003.

[5] Z. Qizhong, "Gene selection and classification using non-linear kernel support

vector machines based on gene expression data," *IEEE/ICME International Conference on Complex Medical Engineering*, 2007.

[6] M. Xiong, X. Fang, and J. Zhao, "Biomarker identification by feature wrappers," *Genome Research*, vol. 11, no. 11, pp. 1878–1887, 2001.

[7] (2012, Sep) Database for annotation, visualization and integrated discovery (david) @ONLINE. http://david.abcc.ncifcrf.gov/.

[8] B. L. Bowerman and R. T. O'Connell, *Time series forecasting: unified concepts and computer implementation*. Boston, MA, USA: PWS Publishing Co., 1986.

[9] E. E. B. Institute. (2012, Oct) What is bioinformatics? @ONLINE. http://www.ebi.ac.uk/2can/bioinformatics/bioinf$_w hat_1 .html$.

[10] N. N. C. Institute. (2012, Oct) What is cancer? @ONLINE. http://www.cancer.gov/cancertopics/cancerlibrary/what-is-cancer.

[11] P. Health. (2012, Oct) Breast cancer @ONLINE. http://www.ncbi.nlm.nih.gov/pubmedhealth/PMH0001911/.

[12] A. Doncescu, "Machine learning applied to brca1 hereditary breast cancer data," *International Conference on Advanced Information Networking and Applications Workshops, 2009. WAINA '09.*, pp. 942 – 947, May 2009.

[13] J. Hall, M. Lee, B. Newmanand, J. Morrow, L. Anderson, B. Huey, and M. King,

"Linkage of early-onset familial breast cancer to chromosome 17q21," *Science*, vol. 250, pp. 1684–1689, 1990.

[14] W. Hofmann and P. Schlag, "Brca1 and brca2-breast cancer susceptibility genes," *Journal of Cancer Research and Clinical Oncology*, vol. 126, pp. 487–496, 2000.

[15] K. Pääkkönen, S. Sauramo, L. Sarantaus, P. Vahteristo, A. Hartikainen, P. Vehmanen, J. Ignatius, V. Ollikainen, H. Kriinen, E. Vauramo, H. Nevanlinna, R. Krahe, K. Holli, and J. Kere, "Involvement of brca1 and brca2 in breast cancer in a western finnish sub-population," *Gentic Epidemiology*, vol. 20, pp. 239–246, 2001.

[16] S. N. Powell and L. A. Kachnic, "Roles of brca1 and brca2 in homologous recombination, dna replication fidelity and the cellular response to ionizing radiation," *Oncogene*, vol. 22, pp. 5784–5791, 2003.

[17] J. M. Satagopan, K. Offit, W. Foulkes, M. E. Robson, S. Wacholder, C. M. Eng, S. E. Karp, and C. B. Begg, "The lifetime risks of breast cancer in ashkenazi jewish carriers of brca1 and brca2 mutations," *Cancer Epidemiol Biomarkers Preview*, vol. 10, pp. 467–473, 2001.

[18] D. Ford, D. Easton, D. Bishop, S. Narod, and D. Goldgar, "Risks of cancer in brca1-mutation carriers. breast cancer linkage consortium," *Lancet*, vol. 343, no. 8899, pp. 692–5, March 1994.

[19] E. Schubert, M. Lee, H. Mefford, R. Argonza, J. Morrow, J. Hull, J. Dann, and M. King, "Brca2 in american families with four or more cases of breast or ovar-

ian cancer: recurrent and novel mutations, variable expression, penetrance, and the possibility of families whose cancer is not attributable to brca1 or brca2," *American Journal of Human Genetics*, vol. 60, no. 5, pp. 1031–1040, 1997.

[20] M. Raza, I. Gondal, D. Green, and R. L. Coppel, "Feature selection and classification of gene expression profile in hereditary breast cancer," in *Proceedings of the Fourth International Conference on Hybrid Intelligent Systems*, 2004.

[21] S. Ma, M. Shi, Y. Li, D. Yi, and B.-C. Shia, "Incorporating gene co-expression network in identification of cancer prognosis markers," *BMC Bioinformatics*, vol. 11, pp. 271–281, 2010.

[22] L. E. Baum, T. Petrie, G. Soules, , and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains," *The Annals of Mathematical Statistics*, vol. 41, no. 1, pp. 164–171, 1970.

[23] T. Fawcett, "Roc graphs: Notes and practical considerations for researchers," *Machine Learning*, vol. 31, pp. 1–38, 2004.

[24] A. P. Bradley, "The use of the area under the roc curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, no. 7, pp. 1145 – 1159, 1997.

[25] A. H. Chen, Y.-W. Tsau, and J. . Ching-Heng Lin, "Novel methods to identify biologically relevant genes for leukemia and prostate cancer from gene expression profiles."

[26] D. K. Slonim, P. Tamayo, J. P. Mesirov, T. R. Golub, and E. S. Lander, "Class prediction and discovery using gene expression data," in *Proceedings of the fourth annual international conference on Computational molecular biology*. ACM, 2000, pp. 263–272.

[27] C. Lazar, J. Taminau, S. Meganck, D. Steenhoff, A. Coletta, C. Molter, V. de Schaetzen, R. Duque, H. Bersini, , and A. Nowé, "A survey on filter techniques for feature selection in gene expression microarray analysis," vol. 9, no. 4, pp. 1106–1119, 2012.

[28] Y. Saeys, I. Inza, and Larra, "A review of feature selection techniques in bioinformatics."

[29] P. Park, M. Pagano, M. Bonetti *et al.*, "A nonparametric scoring algorithm for identifying informative genes from microarray data." in *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 2001, p. 52.

[30] H. Chuang, H. Tsai, Y. Tsai, and C. Kao, "Ranking genes for discriminability on microarray data," 2003.

[31] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. Lander, and T. Golub, "Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation," *Proceedings of the National Academy of Sciences*, vol. 96, no. 6, pp. 2907–2912, 1999.

[32] P. L. Lanzi, "Fast feature selection with genetic algorithms: a filter approach,"

in *Evolutionary Computation, 1997., IEEE International Conference on.* IEEE, 1997, pp. 537–540.

[33] J. Yang and V. Honavar, "Feature subset selection using a genetic algorithm," *Intelligent Systems and Their Applications, IEEE*, vol. 13, no. 2, pp. 44–49, 1998.

[34] W. Siedlecki and J. Sklansky, "A note on genetic algorithms for large-scale feature selection," *Pattern Recognition Letters*, vol. 10, no. 5, pp. 335–347, 1989.

[35] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[36] B. Boser, I. Guyon, and V. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory.* ACM, 1992, pp. 144–152.

[37] T. Furey, N. Cristianini, N. Duffy, D. Bednarski, M. Schummer, and D. Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data," *Bioinformatics*, vol. 16, no. 10, pp. 906–914, 2000.

[38] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 13, pp. 37 – 52, 1987.

[39] T. K. Seize, "Student's t-test," *Southern Medical Journal*, vol. 70, no. 11, p. 1299, 1977.

[40] M. R. Hassan, M. M. Hossain, J. Bailey, G. Macintyre, J. W. Ho, and K. Ramamo-

hanarao, "A voting apporach to indentify a small number of highly predictive genes using multiple classifiers," *BMC Bioinformatics*, 2009.

[41] D. V. Ngueyen and D. M. Rocke, "Multi-class cancer classification via partial least squares with gene expression profiles," *Bioinformatics*, vol. 18, no. 9, pp. 1216–1226, 2002.

[42] Y. Ai-Jun and S. Xin-Yuan, "Bayesian variable selection for disease classification using gene expression data," *Bioinformatics*, vol. 26, no. 2, pp. 215–222, 2009.

[43] X. Zhou, X. Wang, and E. R. Dougherty, "A bayesian approach to nonlinear probit gene selection and classification," *Journal of the Franklin Institute*, vol. 341, no. 1-2, pp. 137–156, 2004.

[44] L. J. v. Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Maron, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Robers, P. S. Linsley, R. Bernards, and S. H. Friend, "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, no. 31, pp. 530–536, 2002.

[45] T. Sørlie, R. Tibshirani, J. Parker, T. Hastie, J. S. Marron, A. Nobel, S. Deng, H. Johnsen, R. Pesich, S. Geisler, J. Demeter, C. M. Perou, P. E. Lnning, P. O. Brown, A.-L. Brresen-Dale, , and D. Botstein, "Repeated observation of breast tumor subtypes in independent gene expression data sets," *Proceedings of the National Academy of Sciences*, vol. 100, no. 14, pp. 8418–8423, 2003.

[46] E. Honrado, A. Osorio, J. Palacios, and J. Benitez, "Pathology and gene expres-

sion of hereditary breast tumors associated with brca1, brca2 and chek2 gene mutations," *Oncogene from Nature*, vol. 25, pp. 5837–5845, 2006.

[47] J. D. Parvin, Z. Kais, M. Arora, S. Kotian, A. Zha, D. Ransburgh, D. Bozdag, U. Catalyurek, and K. Huang, "Identification of a breast cancer associated regulatory network," in *Ohio Collaborative Conference on Bioinformatics*, 2009, pp. 71–75.

[48] S. Kim, E. R. Dougherty, J. Barrera, Y. Chen, M. L. Bittner, and J. M. Trent., "Strong feature sets from small samples," *Journal of Computational biology*, vol. 9, no. 1, p. 127146, 2002.

[49] H. Mamitsuka, "Selecting features in microarray classification using roc curves," *Pattern Recognition*, vol. 39, no. 12, pp. 2393–2404, 2006.

[50] T. G. Tape. (2012, Sep) The area under an roc curve @ONLINE. http://gim.unmc.edu/dxtests/roc3.htm.

[51] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[52] A. Jazaeri, C. Yee, C. Sotiriou, K. Brantley, J. Boyd, and E. Liu, "Gene expression profiles of brca1-linked, brca2-linked, and sporadic ovarian cancers," *Journal of the National Cancer Institute*, vol. 94, no. 13, pp. 990–1000, 2002.

[53] P. Health. (2013, Jan) Home - pubmed @ONLINE.

[54] M. Mamtani and H. Kulkarni, "Association of hadha expression with the risk of

breast cancer: targeted subset analysis and meta-analysis of microarray data," *BMC Research Notes*, vol. 5:25, 2012.

[55] H. M. M̈uller, L. Ivarsson, H. Schrcksnadel, H. Fiegl, A. Widschwendter, G. Goebel, S. Kilga-Nogler, H. Philadelphy, W. Gtter, C. Marth, and M. Widschwendter, "Dna methylation canges in sera of women in early pregnancy are similar to those in advanced breast cancer patients," *Clinical Chemistry*, vol. 50, no. 6, pp. 1065–1068, 2004.

[56] N. S. Fearnhead, M. P. Britton, and W. F. Bodmer, "The abc of apc," *Human Molecular Genetics*, vol. 10, no. 7, pp. 721–733, 2001.

[57] S. O'Malley, H. Su, T. Zhang, C. Ng, H. Ge, and C. K. Tang, "Tob suppresses breast cancer tumorigenesis," *International Jornal of Cancer: Cancer Cell Biology*, vol. 125, no. 8, pp. 1805–1813, October 2009.

[58] M. W. Helms, D. Kemming, C. H. Contag, H. Pospisil, K. Bartkowiak, A. Wang, S.-Y. Chang, H. Buerger, and B. H. Brandt, "Tob1 is regulated by egf-dependent her2 and egfr signaling, is highly phosphorylated, and indicates poor prognosis in node-negative breast cancer," *American Association for Cancer Research*, vol. 69, pp. 5049–5056, 2009.

[59] M. Yano, K. Hirai, Z. Naito, M. Yokoyama, T. Ishiwata, Y. Shiraki, M. Inokuchi, and G. Asano, "Expression of cathepsin b and cystatin c in human breast cancer," *Surgery Today*, vol. 31, pp. 385–389, 2001.

[60] N. Vigneswaran, J. Wu, S. Muller, W. Zacharias, S. Narendran, and L. Middleton,

"Expression analysis of cystatin c and m in laser-capture microdissectioned human breast cancer cells-a preliminary study," *Pathology - Research and Practice*, vol. 200, pp. 753–762, 2005.

[61] F. M. Tumminello, C. Flandina, M. Crescimanno, and G. Leto, "Circulating cathepsin k and cystatin c in patients with cancer related bone disease: clinical and therapeutic implications," *Biomedicine and Pharmacotherapy*, vol. 62, no. 2, p. 130:5, 2008.

[62] C. Ambrosino, R. Tarallo, A. Bamundo, D. Cuomo, G. Franci, G. Nassa, O. Paris, M. Ravo, A. Giovane, N. Zambrano, T. Lepikhova, O. A. Janne, M. Baumann, T. A. Nyman, L. Cicatiello, and A. Weisz, "Identification of a hormone-regulated dynamic nuclear actin network associated with estrogen receptor α in human breast cancer cell nuclei," *Molecular & Cellular Proteomics*, vol. 9, no. 6, pp. 1353–1367, 2010.

[63] K. Majidzadeh-A, R. Esmaeili, and N. Abdoli, "Tfrc and actb as the best reference genes to quantify urokinase plasminogen activator in breast cancer," *BMC Research Notes*, vol. 4215, June 2011.

[64] M. J. Duffy, "The urokinase plasminogen activator: role in malignancy," *Current Pharmaceutical Design*, vol. 10, no. 1, pp. 39–49, 2004.

[65] E. Favaro, A. Ramachandran, R. McCormick, H. Gee, C. Blancher, M. Crosby, C. Devlin, C. Blick, F. Buffa, J.-L. Li, B. Vojnovic, R. P. das Neves, P. Glazer, F. Inorra, M. Ivan, J. Ragoussis, and A. L. Harris, "Microrna-210 regulates mi-

tochondrial free radical response to hypoxia and krebs cycle in cancer cells by targeting iron sulfur cluster protein iscu," *PLoS ONE*, vol. 5, p. e10345, 2010.

[66] M. Udler, A.-T. Maia, A. Cebrian, C. Brown, D. Greenberg, M. Shah, C. Caldas, A. Dunning, D. Easton, B. Ponder, and P. Pharoah, "Common germline genetic variation in antioxidant defense genes and survival after diagnosis of breast cancer," *Journal of Clinical Oncology*, vol. 25, no. 21, pp. 3015–3023, 2007.

[67] I. Heirman, D. Ginneberge, R. Brigelius-Flohe, N. Hendrickx, P. Agostinis, P. Brouckaert, P. Rottiers, and J. Grooten, "Blocking tumor cell eicosanoid synthesis by gpx4 impedes tumor growth and malignancy," *Free Radical Biology & Medicine*, vol. 40, pp. 285–294, 2006.

[68] D. J. Sugarbaker, W. G. Richards, G. J. Gordon, L. Dong, A. D. Rienzo, G. Maulik, J. N. Glickman, L. R. Chirieac, M.-L. Hartman, B. E. Taillon, L. Du, P. Bouffard, S. F. Kingsmore, N. A. Miller, A. D. Farmer, R. V. Jensen, S. R. Gullans, and R. Bueno, "Transcriptome sequencing of malignant pleural mesothelioma tumors," *PNAS*, vol. 105, no. 9, pp. 3521–3526, 2008.

[69] I. P. P. A. . Reagents. (2012, Oct) Transcriptional factors and regulators @ONLINE. http://www.imgenex.com/TranscriptionFactors.php.

[70] P. Broun, Y. Liu, E. Queen, Y. Schwarz, M. L. Abenes, and M. Leibman, "Importance of transcription factors in the regulation of plant secondary metabolism and their relevance to the control of terpenoid accumulation," *Phytochemistry Reviews*, vol. 5, no. 1, pp. 27–38, 2006.

[71] (2012, Sep) Transfac @ONLINE. http://www.gene-regulation.com/pub/databases/transfac/doc/toc.html.

[72] V. V. Vazirani, *Approximation Algorithms*, 1st ed. Springer, 2001.

[73] N. C. for Biotechnology Information. (2012, Sep) Atf6 activating transcription factor 6 [homo sapiens] @ONLINE. http://www.ncbi.nlm.nih.gov/gene/22926.

[74] M. Shuda, N. Kondoh, N. Imazeki, K. Tanaka, T. Okada, K. Mori, A. Hada, M. Arai, T. Wakatsuki, O. Matsubara, N. Yamamoto, and M. Yamamoto, "Activation of the atf6, xbp1 and grp78 genes in human hepatocellular carcinoma: a possible involvement of the er stress pathway in hepatocarcinogenesis," *Journal of Hepatology*, vol. 38, no. 5, pp. 605–614, 2003.

[75] Y. Mao, X. Zhou, D. Pi, Y. Sun, , and S. T. C. Wong, "Multiclass cancer classification by using fuzzy support vector machine and binary decision tree with gene selection," *Journal of Biomedicine and Biotechnology*, vol. 2005, no. 2, pp. 160–171, 2005.

Vitae

- Name: Imran ul Haq
- Nationality: Pakistani
- Date of Birth: 22-May-1983
- Email: *immithegreat@gmail.com, immithegreat@hotmail.com*
- Telephone No.: 03-894-9840
- Mobile No.: 056-347-4820
- Present Address: P. O. Box 8616, KFUPM, Dhahran 31261, KSA
- Permanent Address: House # 11/5, Block 2K, Nazimabad No. 2, Karachi, Pakistan
- Last Degree: Bachelor of Science in Computer Engineering
- Institute: Sir Syed University of Engineering and Technology, Karachi
- GPA: 3.27/4.0
- Year Completed: 2008